Data driven construction of MHD surrogates using sparse regression and data assimilation

Alasdair Roy

Submitted in accordance with the requirements for the degree of

Doctor of Philosophy and Master of Science

University of Leeds

Centre of Doctoral Training in Fluid Dynamics

School of Computing

June 2025

I confirm that the work submitted is my own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgments

I would like to begin by thanking my supervisors Steve Tobias, Phil Livermore and Wayne Arter for their dedication and patience with this project. Special thanks is owed to Wayne for helping me during my time at UKAEA. I am also grateful to Chris Jones and Stan Pamela for their contributions when they were involved with my project. I thank the CDT for support throughout my degree and to UKRI for sponsoring this project (grant no. EP/S022732/1). This work has been (part-) funded by the EPSRC Energy Programme [grant number EP/W006839/1].

I am grateful to the many friends I made throughout my time in the CDT, and their own personal contributions to my life through coffee breaks, climbing and many weekends spent in the office rather than outside. Without them all, I doubt I would have made it this far.

Having now spent five years in Leeds, I owe substantial thanks to the many friends I made outside of my PhD. I thank Paul and Ellie for their support, Tyler for providing me with much needed company while writing the thesis, Zack and Julia for supporting me throughout my studies, and Ioana for taking me outdoors climbing during the times it was most needed. I also thank my parents for their continued support during my extensive time in academia.

Abstract

Tokamak operation is plagued by the presence of magnetohydrodyamic instabilities which impose limitations on their efficiency and can cause early termination of the plasma. Understanding of many of these instabilities comes from nonlinear numerical simulations of resistive magnetohydrodynamics which are challenging to perform owing partly to the timescales that must be resolved. Fortunately, simplified ordinary differential equations called the ANAC and ANAET models can be derived using symmetry arguments with bifurcation theory which display qualitative similarities to observed tokamak instabilities. The qualitative similarity of these models motivates exploring approaches which allow them to be related quantitatively to experiment.

In this dissertation we implement two data-driven approaches which can be used to either derive simplified models of tokamak instabilities or be used to match already known simplified models to experimental diagnostics. The first of these methods is a popular regression framework called the sparse identification of nonlinear which we validate on a low-dimensional model of magnetoconvection behaviour and use to derive low-dimensional models directly from numerically simulated magnetoconvection PDE data. We suggest that implementation of the weak form and constraints are almost certainly required in future applications. Results show that models derived from POD modes of magnetoconvection PDE data can show expected bifurcations present in the PDE.

The second approach is called the ensemble Kalman filter and is applied to two models which resemble the sawtooth instability in tokamaks. We demonstrate how the ensemble Kalman filter can be used for parameter estimation of these two models in experiment like conditions, displaying robustness to high degrees of noise, low sampling rates and multiscale dynamics. By using a stochastic integration scheme, we draw parallels between observed sawtooth instabilities in tokamaks and the ANAET model.

Contents

List of Figures

List of Tables

1 Introduction

In 2022, the world population reached 8 billion for the first time and is set to continue to rise until peaking in the 2080s at over 10 billion [176]. During this time, global energy demand will increase substantially and meeting these needs will become a serious challenge. Currently most of the world's energy production is met with fossil fuels with their consumption still increasing [143]. The reliance on fossil fuels to fill this increase in energy demand poses serious issues. The first issue is that they constitute a finite resource and will eventually run out. The second larger issue is their contribution to global warming and increase in atmospheric CO_2 which has resulted in an increase in global temperature [78].

Despite the Paris agreement, CO_2 emissions have not yet peaked [144] and predictions based on current policies expect that we will surpass the threshold of 1.5°C above pre-industrial levels [177, 178]. This has placed an ever-growing emphasis on the need to generate energy from renewables or non-carbon based energy sources. Within non-carbon sources, nuclear fission is currently the most viable at present. In the world there are around 440 nuclear power plants, however, due to negative public opinion new reactors have rarely been built with the last power plant in the UK constructed one year after the Chernobyl accident in 1986. This event has skewed the perception of nuclear despite nuclear having an otherwise exceptional safety record compared to fossil fuels [58][pg. 10]. Another major concern with nuclear fission plants is the long-lived radioactive waste, an issue which would have to be addressed. Despite these perceptions new sites are set to open to meet increasing energy needs [181].

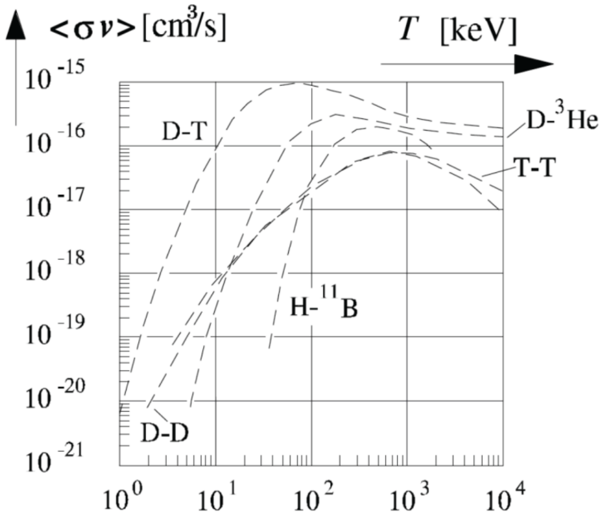


Figure 1: Reaction cross-sections denoted by $\langle\sigma v\rangle$ for some fusion reactions with temperature. The most favourable reaction is given by D-T. Figure taken from [78].

An appealing alternative to fission is nuclear fusion. As opposed to splitting apart heavy atoms to produce neutrons in a self-sustaining reaction fusion binds together light atomic nuclei producing minimal radioactive waste of different character to fission. Fusion has many advantages: there are no direct CO_2 emissions, fuel sources have a long-time supply, and the radioactive waste produced is short-lived compared to fission [58][pg. 17]. There are two main approaches for nuclear fusion, namely inertial confinement fusion and magnetic confinement fusion [88][§7 & §9]. The former typically uses lasers to cause ablation of the surface of a fuel pellet, causing it to implode and thus fusion to take place. The latter confines the fuel using magnetic fields at high temperatures. Unfortunately, economic generation of electricity from fusion through either approach turns out to be extremely complex, making fusion a long-term solution to energy generation problems [58][pg. 6].

The most favourable reaction is given by the fusion of the hydrogen isotopes called deuterium (one neutron) and tritium (two neutrons)

$$D + T \rightarrow ^4 He\,(3.5MeV) + n\,(14.1)\,MeV \tag{1.1}$$

where the energy results from a mass-deficit in the reaction and we use D to denote deuterium and T for tritium. The reaction produces a neutron (n) and a helium nucleus (alpha particle). For the reaction to take place the two nuclei must collide and overcome the Coulomb repulsion force. At low energies the reaction has a small cross-section (essentially the probability of reaction). Figure 1 shows the cross-section of some plausible reactions highlighting that the D-T reaction has higher cross-sections at lower energies. Classically for this reaction to take place we would need temperatures more than 3 billion Kelvin ($1keV \approx 11600K$), however due to quantum tunneling and other cross-sectional dependencies on temperature, lower temperatures can be used [19][pg. 5]. In reality, temperatures of around 200 million Kelvin are required for particles to have enough energy for fusion to occur [88][pg. 43].

Deuterium is an isotope of hydrogen containing one neutron and can be found abundantly in water. On the other hand, tritium is a radioactive isotope of hydrogen which is not found abundantly in nature due to its short half-life [58][pg. 27]. However, it can be obtained from lithium

$$^6Li \ + \ n \ \rightarrow ^4 He \ (2.05MeV) \ + \ T \ (2.73MeV) \tag{1.2}$$

and so neutrons from the fusion reaction are used to bombard Lithium breeding blankets and produce tritium. Lithium is a relatively abundant resource and so the consumable fuels in fusion are really Li and D [88][pg. 137].

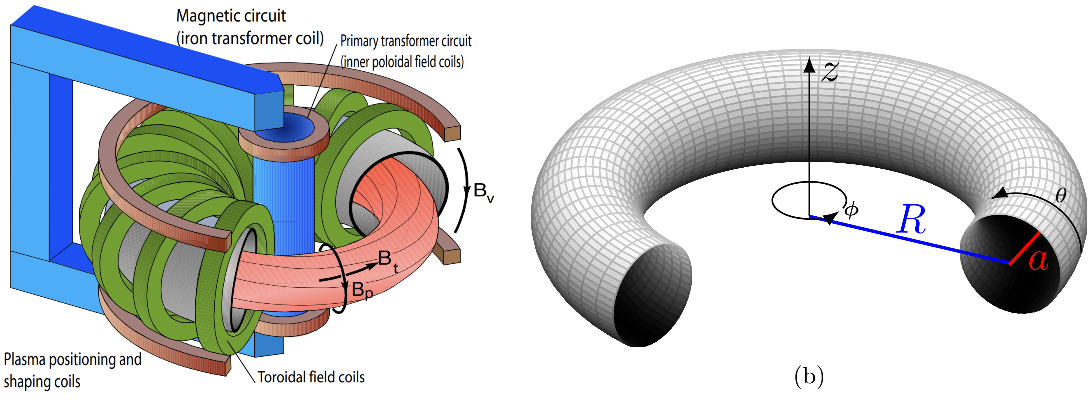
One of the main objectives in fusion is creating net energy output. To create the conditions for fusion, energy must be supplied to the reactor and eventually we would require that the energy out exceeds the energy supplied. In the D-T reaction around 80% of the energy is carried by the neutron and the release of this neutron is how fusion reactors generate energy for thermal heating. The alpha particle carries around 20% of the energy and is retained in magnetic confinement devices owing to its charge and this can be used to sustain the fusion reaction. At sufficiently high temperatures, alpha particle heating alone can be enough to sustain the fusion reaction and this point is called ignition. The ignition criterion is calculated by finding when alpha particle heating balances the losses [88]. The condition for ignition is

$$n \times \tau_E > 1.7 \times 10^{20} m^{-3} s, \tag{1.3}$$

where n is the plasma density and τ_E is the energy confinement time. The energy confinement time expresses the ratio of total plasma energy to the rate at which energy is lost from the plasma [88][pg. 41]. In D-T reactions this corresponds to energies of around 30keV. However, as cross-sections and other parameters depend on plasma temperature, temperatures of 10-20keV are favourable [88][pg. 43]. This corresponds to temperatures of 100-200 million Kelvin, hotter than the surface of the Sun. The ignition condition is more commonly written as the fusion-triple product [88][pg. 43]

$$nT\tau_E = 3 \times 10^{21} m^{-3} keV s, \tag{1.4}$$

where T is the plasma temperature.



(a)

(b)

Figure 2: a) Schematic diagram of a tokamak showing the general configuration of the magnetic field coils and plasma. Figure taken from [90]. b) Circular cross-section toroidal co-ordinate system with minor radius a and major radius R.

Another view of a viable reactor is the case where a self-sustained reaction is not achieved but the energy return exceeds the energy input. This is expressed as the ratio

$$Q \equiv \frac{\text{fusion power produced}}{\text{heating power supplied}} \tag{1.5}$$

where the ignited condition corresponds to $Q \to \infty$. A Q value of 5 or more has energy yield comparable to the external heating, but still remains far from economic energy production [42][pg. 6].

At fusion temperatures the fuel is fully ionized and there are no materials which can withstand direct contact. A magnetic field must then be used to confine the plasma away from the reactor walls. This occurs because of the Lorentz force

$$F = q_e \left(\boldsymbol{v} \times \boldsymbol{B} \right), \tag{1.6}$$

where q_e is the particle charge, $\boldsymbol{v}$ the particle velocity and $\boldsymbol{B}$ the magnetic field. Immediately it can be seen that for a particle of charge q_e no force is exerted if they travel parallel to the magnetic field. Forces are only exerted on motions perpendicular to the magnetic field and as a result particles gyrate around magnetic field lines.

Early fusion devices were often linear in design resembling long solenoids with an applied axial field. Particles were therefore confined to travel along the axis of the device, however, end losses must be dealt with. One solution are so-called magnetic mirrors at each end which reflect particles through an increasing magnetic gradient. Despite this, end losses in these devices were still too high [112][§10.2.1.2]. A natural solution to end losses is to join the two ends of the containment device in a torus shape, thus closing magnetic field lines on themselves in the toroidal direction (the long way around the torus) with general geometry shown in Figure 2.

It turns out that a toroidal field alone is insufficient for confinement of the plasma in a torus. The magnetic field in a torus with purely toroidal field is necessarily non-uniform and the magnitude of the magnetic field varies inversely with major radius [42][pg. 10]. On the inside of the device closer to the major axis field lines are more crowded, causing the radius of gyration to be larger on one side of each orbit than the other [112][pg. 370]. This gives rise to what is called a grad-B vertical drift for electrons and ions in opposite directions. This separation of charge in turn results in a vertical electric field, producing an $\boldsymbol{E} \times \boldsymbol{B}$ drift which is the same for both particles and causes a drift radially outward losing confinement shown in Figure 3. It is therefore necessary to include a poloidal field component (magnetic field the short way around the torus) so that the resulting field winds helically around the torus. The poloidal field comes from the current in the plasma itself which is generated by transformer action shown schematically in Figure 2 a). A current is passed through the inner poloidal field coils which creates a flux change in the torus inducing a current in the toroidal direction [16]. As the current is generated by transformer action tokamaks are pulsed devices and do not operate indefinitely.

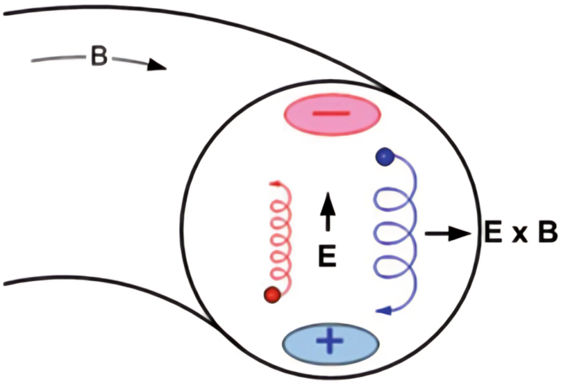


Figure 3: $\boldsymbol{E} \times \boldsymbol{B}$ resulting from a purely toroidal field in a torus. Figure adapted from [112].

The plasma current also serves the second important function of heating the plasma through plasma collisions. This is called Ohmic heating, but becomes less effective at heating the plasma as it becomes hotter. For this reason, plasmas can only reach up to 50 million degrees when heated Ohmically which still falls short of the requirement for ignition [88]. Secondary heating methods are therefore employed within tokamaks to reach higher temperatures. One common method is neutral-beam injection (NBI) where deuterium ions are accelerated to high energies and heat the plasma by collisions [112][§10]. Before the ions enter the plasma, they must be made neutral otherwise they will be affected by the magnetic field. A consequence of NBI is that the momentum transfer can cause the plasma to rotate in the laboratory frame of reference.

During research of the tokamak it was experimentally observed that an increase in plasma current resulted in an improvement of both confinement times and plasma temperature [88][§10.1]. This prompted the design of the Joint European Torus (JET) which began operations in 1983. JET showed significantly improved confinement times versus previous tokamaks, able to reach higher plasma temperatures than previously achieved. In 1997, during a series of D-T campaigns, JET achieved a Q value of $Q \simeq 0.67$ but has since reached the end of its operational life-span in December 2023. Since the 2000s, much of JET's operations have been in support of the design of a new tokamak named the International Thermonuclear Experimental Reactor (ITER) in Cadarache in the south of France. ITER is a joint international collaboration which is set to achieve Q values of $Q \geq 10$ with pulse durations of $t = 400s$, demonstrating a hot plasma mainly heated by alpha particles [95]. Successful operation of ITER is expected to lead to the development of the first commercially viable reactor known as DEMO.

A significant amount of work has been dedicated to the developmental challenges remaining for ITER outlined on the Eurofusion page [180]. One of the primary concerns is the role that instabilities will play in ITER's operations. For example, the on axis current which is used to both heat the tokamak and generate the poloidal field is also the driving source of prominent large scale magnetohydrodynamic (MHD) instabilities [54]. One such instability is the kink mode which is thought to be responsible for the sawtooth instability which produces a periodic rise of the central plasma temperature followed by a sudden crash and re-organisation of the magnetic field [34]. While the instability does not necessarily inhibit tokamak operation, other instabilities certainly can result in termination of the plasma and ultimately damage plasma facing components. There is also further concern that the sawtooth instability can seed other more deleterious instabilities which will cause termination of the plasma and loss of confinement, thus limiting the achievable performance [34]. Resistive and ideal MHD play a key role in understanding the cause of such instabilities [139], but what remains clear is that performing such simulations is challenging [131], with the exact mechanism behind instabilities in many cases remaining disputed [136]. The complex nature of these instabilities underpins the need to build simplified models of their behaviour.

The prominence of MHD instabilities during operation is the result of many tokamaks operating near instability limits because this is where the best performance is achieved [56]. This observation has led to the application of what is called bifurcation theory [71, 72], which performs a qualitative study how solution behaviours change as parameters are varied. Close to the onset of instability, we assume that the system is governed by a small number of ordinary-differential equations that topologically capture the behaviour of the full system. Bifurcation theory can be viewed as a generalization of linear stability theory, where higher order terms in the Taylor series expansion of say partial differential equation variables are retained and nonlinear models are developed. The derivation of these models can be further constrained by the symmetries present in a tokamak, namely an invariance of the solutions in the toroidal and poloidal directions or constrained by energy preservation where ideal magnetohydrodynamics is relevant [72]. However, modern tokamaks typically feature non-circular cross sections due to plasma shaping and divertors which aid with plasma exhaust [16][§1.6], so the poloidal symmetry is doubtful. In any case, a number of models have been derived which stand as candidates for observed tokamak instabilities [71] and provide at least qualitative information on instability behaviour.

The next question that can be asked is how these qualitative models can be related quantitatively to either experimental diagnostics or numerical models. If models do perform successful prediction of experiment or numerics, they could yield either methods for control or some insight into the underlying physics [72]. Equating these models to experiment is, however, far from simple. Many instabilities are often on the order of millisecond or faster events [16][§7.1], separated by longer quiescent periods [131]. This behaviour also makes accurate measurement and diagnosis of instabilities a hard experimental challenge and for this reason tokamaks have a wide range of diagnostics [68]. Tokamaks are also clearly hostile environments where large plasma temperatures and neutron fluxes make diagnosing the plasma a major challenge. Observations used to diagnose the sawtooth instability for example are often both noisy and poorly sampled in time [26, 166]. Fortunately, many of the diagnostics have improved through the years [166] and there are now an abundance of time series analysis techniques that can be used [48, 97, 155].

In this dissertation, we consider two approaches for relating models of the type listed in ref. [71] and ref. [72] to either experimental diagnostics or numerics. The first of these approaches is called the sparse identification of nonlinear dynamics, or SINDy for short [97]. SINDy has emerged as a popular technique which takes a set of input time series and performs model selection from a candidate feature library, returning a sparse model (for example a set of ordinary differential equations) which reproduce the dynamics. The interest in SINDy has been driven by the notion of producing generalisable, interpretable models, stepping away from other black-box approaches. The second approach we consider is called the ensemble Kalman filter and falls under the category of data assimilation [24]. Data assimilation methods have a long history in application with weather prediction, and are noted to be noise robust making the approach an appealing candidate [69]. Both of these techniques will be explored in detail.

We begin by discussing the technical context and background of the project given in §2, with a particular focus on the derivation of simplified bifurcation theory models from ref. [72]. We conclude this section by offering a comparison between experimental data and the simplified models to motivate their similarity. In §3.1 we introduce SINDy and review the literature relevant to this dissertation, with a focus on noise robustness and conceptual challenges. This is followed by a benchmarking of SINDy in §4 to a set of ordinary differential equations derived from magnetohydrodynamics. Following benchmarking, we then discuss application of SINDy to time series derived from the numerical simulation of partial differential equations in §5, discussing how SINDy can be used to find simplified models of high-dimensional systems. In §6 we review the ensemble Kalman filter, its extensions, and the implementation of the methods discussed in the review as a Python code. The software is then validated in several cases to address concerns which have been raised in previous works [106]. The ensemble Kalman filter is then extended in §7 to a more complicated scenario in which training data is generated from a surrogate model which closely resembles experimental data. We present a progressive study of several cases which intend to leave the ensemble Kalman filter in a position to be applied successfully to experimental data. We compare the performance of the ensemble Kalman filter with SINDy in §8 and discuss the comparison of the two approaches when applied to experimental data. The conclusions of this dissertation and avenues for future research are presented in §9.

2 MHD, tokamak stability and equivariant bifurcation theory models

In this chapter we discuss the overall context for this dissertation. The chapter begins by briefly introducing the MHD equations which are widely applied in understanding tokamak instabilities. We then introduce some MHD tokamak instabilities with a particular focus on their qualitative behaviours. Following this, derivations of equivariant bifurcation theory models which describe their behaviour are introduced. The relation between these models and observations of the instabilities on tokamak diagnostics is discussed qualitatively.

2.1 Magnetohydrodynamics

We begin by stating the governing equations for resistive MHD, a derivation of which can be found in many textbooks [41] [pg 58]. MHD describes the plasma in a single-fluid approximation with one governing set of equations for both electrons and ions. The equations are

$$\frac{D\rho}{Dt} = -\rho\nabla\cdot\boldsymbol{u}, \qquad \text{Mass Conservation} \qquad (2.1)$$

$$\rho\frac{D\boldsymbol{u}}{Dt} = (\nabla\times\boldsymbol{B})\times\boldsymbol{B}/\mu_0 - \nabla P + \mu_f\nabla^2\boldsymbol{u}, \qquad \text{Momentum balance} \qquad (2.2)$$

$$\frac{DP}{Dt} = -\gamma P\nabla\cdot\boldsymbol{u} + (\gamma-1)\frac{(\nabla\times\boldsymbol{B})^2}{\sigma\mu_0^2}, \qquad \text{Energy Equation} \qquad (2.3)$$

$$\frac{\partial\boldsymbol{B}}{\partial t} = \frac{\nabla^2\boldsymbol{B}}{\sigma\mu_0} + \nabla\times(\boldsymbol{u}\times\boldsymbol{B}), \qquad \text{Induction equation} \qquad (2.4)$$

where ρ is the plasma density, $\boldsymbol{u}$ the plasma velocity, t is time, $\boldsymbol{B}$ the magnetic field, P the plasma pressure, γ is ratio of specific heats, μ_f is the fluid viscosity, σ is the electrical conductivity and μ_0 is the permeability of free space. We further define the equation of state for the plasma temperature T

$$T = \frac{P}{R_0\rho} \qquad (2.5)$$

where R_0 is the ideal gas constant. For magnetic fields we also have the solenoidal constraint

$$\nabla\cdot\boldsymbol{B} = 0. \qquad (2.6)$$

We also have two definitions

$$\boldsymbol{E} = \frac{(\nabla\times\boldsymbol{B})}{\sigma\mu_0} - \boldsymbol{u}\times\boldsymbol{B} \qquad \text{Ohm's Law,} \qquad (2.7)$$

$$\boldsymbol{j} = \frac{(\nabla\times\boldsymbol{B})}{\mu_0} \qquad \text{Ampere's Law} \qquad (2.8)$$

which assume a non-relativistic plasma and $\boldsymbol{j}$ is the current density and $\boldsymbol{E}$ the electric field. In tokamak applications, a lot of understanding can be gained by neglecting all dissipation in the resistive equations and considering what are known as the *ideal* MHD equations. To understand when this is possible, we look at the balance of the terms in Equation (2.4)

$$\frac{\sigma\mu_0|\nabla\times(\boldsymbol{u}\times\boldsymbol{B})|}{|\nabla^2\boldsymbol{B}|} \sim \frac{\mu_0\sigma L_H^2}{\tau_H} \equiv \mathcal{R}_m \qquad (2.9)$$

where $\mathcal{R}_m$ is the magnetic Reynolds number. The case where diffusion is negligible is when $\mathcal{R}_m \to \infty$. Similarly by balancing force terms in Equation (2.2) we can write

$$\frac{\rho|D\boldsymbol{u}/dt|}{\mu_f|\nabla^2\boldsymbol{u}|} \sim \frac{\rho L_H^2}{\mu_f \tau_H} \equiv \mathcal{R} \tag{2.10}$$

where $\mathcal{R}$ is the Reynolds number, τ_H is the hydrodynamic time and L_H the hydrodynamic length. We can see that from these balances, neglecting all dissipation in the large scale limit where $L_H \to \infty$. Similarly, by dimensional analysis of the terms in the Equation (2.3), and applying the momentum equation (2.2), we arrive at the adiabatic gas law

$$\frac{1}{\rho^\gamma}\frac{D}{Dt}(P\rho^{-\gamma}) = 0 \implies P\rho^{-\gamma} = \text{const.} \tag{2.11}$$

One important consequence of the ideal MHD equations is known as Alfvén's frozen flux theorem which states that field lines are constrained to move with the fluid. It can be shown from the induction equation and Ohm's law that

$$\frac{D}{Dt}\int_S \boldsymbol{B}\cdot \mathrm{d}\boldsymbol{S} = \int_S \left(\frac{\partial \boldsymbol{B}}{\partial t} - \nabla\times(\boldsymbol{u}\times\boldsymbol{B})\right)\cdot \mathrm{d}\boldsymbol{S} = 0 \tag{2.12}$$

for a surface S moving with the fluid. This implies for any surface moving with the fluid closed by a bounded contour C, the total flux passing through that surface is conserved in ideal MHD.

In most plasmas, the magnetic Reynolds number is typically very large $\sim 10^8$, and so it might seem reasonable to neglect resistivity altogether. While we describe ideal MHD as the limit where $L_H \to \infty$, there may exist regions in the plasma where there are shortening length scales of interest, such as near boundaries. Such behaviour is relevant when narrow current sheets form with the plasma and in these regions resistivity becomes increasingly important. Finite resistivity allows the field lines to 'slip' free from the plasma and reconnect to reach lower energy configurations in the plasma and is relevant in what are known as tearing mode instabilities. This process is shown schematically in Figure 4 where oppositely directing field lines form a region with zero magnetic field along the neutral axis. In this region, a narrow current sheet forms and the length scale of interest is then significantly shorter. Finite resisitivity becomes important and magnetic reconnection causes the field lines to change topology. Tension in the magnetic field lines tends to pull magnetic field lines away from the point of reconnection, resulting in the formation of magnetic islands with lower magnetic potential energies. While reconnection may take place in narrow current sheets, the changes to topology of the plasma can affect much larger scales of the system.

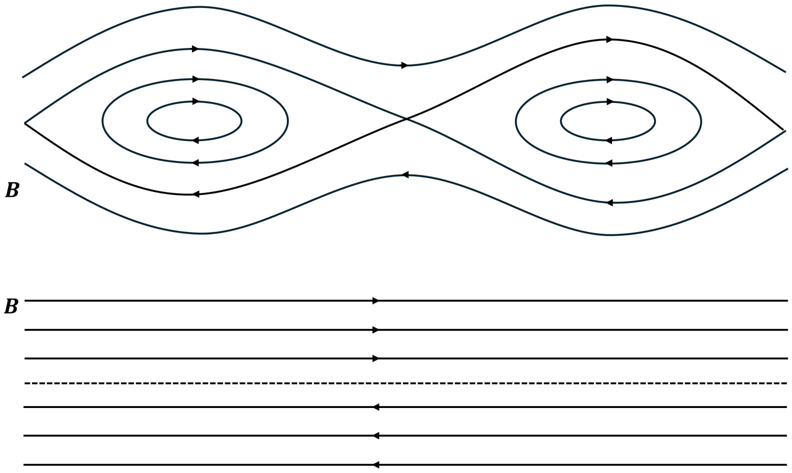


Figure 4: Reconnection of magnetic field lines in resistive MHD leading to the formation of magnetic islands and “X-points”.

2.1.1 The Lorentz force

The momentum equation (2.2) includes the Lorentz force which describes the force exerted by the magnetic field on the fluid given by

$$\boldsymbol{F}_{\mathrm{Lor}} = \boldsymbol{j} \times \boldsymbol{B} = \mu_0^{-1}(\nabla \times \boldsymbol{B}) \times \boldsymbol{B}. \tag{2.13}$$

An intuitive understanding of this force term can be understood by expressing it in terms of components tangential and normal to the magnetic field. We introduce the unit vector $\boldsymbol{b} = \boldsymbol{B}/|\boldsymbol{B}|$, as well as the gradient perpendicular to the magnetic field

$$\nabla_{\perp} = \nabla - \nabla_{||}$$

and the parallel components

$$\nabla_{||} = \boldsymbol{b} \cdot \nabla.$$

The Lorentz force can be written with these definitions as [42][pg. 85]

$$\boldsymbol{j} \times \boldsymbol{B} = \underbrace{\mu_0^{-1} B^2 \boldsymbol{b} \cdot \nabla \boldsymbol{b}}_{\text{magnetic tension}} - \underbrace{\mu_0^{-1} B \nabla_{\perp} B}_{\text{perp. magnetic pressure}}. \tag{2.14}$$

The Lorentz force has been decomposed in terms of a magnetic tensile force which acts to straighten field lines when they are bent and a magnetic pressure terms which acts perpendicular to the the magnetic field and resists compression of the field. This helps guide the basic principle of how we may begin to confine a plasma.

2.1.2 The safety factor and plasma beta

We now introduce two key parameters used in describing toroidal confinement devices which will be referred to when discussing tokamak instabilities. These are called the plasma beta β and the safety factor q, both playing an important role in plasma stability [42][pg. 12]. In an equilibrium state in a tokamak, the field lines wind around nested surfaces in helical paths. The safety factor measures the pitch of the field lines as they traverse the tokamak. q can be measured by the toroidal distance $\delta\phi$ travelled in the time it takes for the field to complete one poloidal rotation

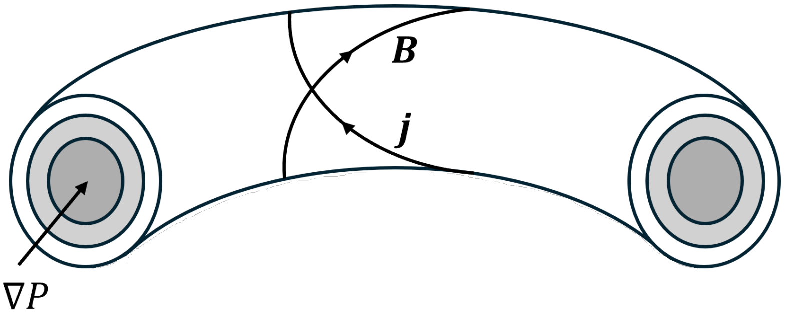


Figure 5: Nested flux surfaces of constant pressure in a cylindrical tokamak. $\boldsymbol{B}$ and $\boldsymbol{j}$ lie on these surfaces of constants pressure under an ideal static force balance.

[16]

$$q = \frac{\delta \phi}{2\pi}. \tag{2.15}$$

It can be seen that small values of q correspond to a tightly wound helix and higher values of q less tightly wound. For a specific surface $r < a$ in a tokamak with circular cross-section, the safety factor is also expressed as $q(r) = aB_\phi / R_0 B_\theta$. Rational values of q play an important role in determining plasma stability [112][pg. 373]. If a perturbation has a wavelength with toroidal mode number m and poloidal mode number n, then it is resonant on the magnetic flux surface where

$$q(r) = \frac{m}{n} \tag{2.16}$$

for $m, n \in \mathbb{Z}$ [42][§3.3]. In tokamak static equilibirum, $q(r)$ increases monotonically from the centre and the variation of q is called magnetic shear [42][pg. 12]. The plasma beta gives a measure of the efficiency of magnetic confinement

$$\beta \equiv \frac{p}{B^2/2\mu_0}, \tag{2.17}$$

describing the ratio of plasma pressure to magnetic pressure [19][§3.5].

2.1.3 Static equilibria

Basic confinement strategies in a tokamak can be understood at a high level from force balances in ideal MHD. A direct consequence of the frozen flux theorem is that by controlling the magnetic field, confinement of an ideal plasma is possible as the field is fixed to the fluid. We begin by considering what are known as equilibrium configurations where the fluid is at rest. Again we require that the diffusion time τ_{diff} is greater than any other time of interest. That is we assume that other instabilities will arise before diffusion becomes relevant. By setting the velocity to zero in the ideal MHD equations, we obtain the static balance MHD equations

$$\boldsymbol{j} \times \boldsymbol{B} = \nabla P, \tag{2.18}$$

$$\nabla \cdot \boldsymbol{B} = 0, \tag{2.19}$$

$$\boldsymbol{j} = \mu_0^{-1} \nabla \times \boldsymbol{B}. \tag{2.20}$$

For a confinement equilibrium, we thus require that the fluid pressure balances the Lorentz force. It also follows that

$$\boldsymbol{B} \cdot \nabla P = \boldsymbol{j} \cdot \nabla P = 0 \tag{2.21}$$

and so both $\boldsymbol{j}$ and $\boldsymbol{B}$ lie on surfaces of constant pressure with normal defined by ∇P, as shown in Figure 5. Simply put, the Lorentz force acts opposite to the pressure toward the central axis of the plasma and so the plasma can be contained through magnetic confinement.

2.1.4 Pinch fields

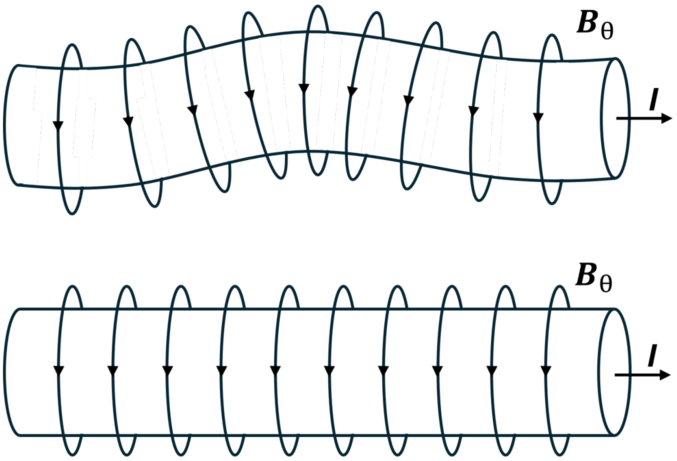


Figure 6: The kink instability causes radial displacement of a plasma column (top).

Early fusion experiments relied on this equilibria condition to confine plasmas [112][§10.2.1.1]. First consider a cylindrical polar system (r, θ, z) where the current density $\boldsymbol{j} = j\hat{\boldsymbol{z}}$ and $\boldsymbol{B} = B(r)\hat{\boldsymbol{\theta}}$. This configuration is known as the $z-$pinch and was used for early plasma confinement. However, it turns out that this configuration is highly unstable to the kink instability. This perturbation forms a kink in the plasma column which results in the field lines being compressed on one side of the column, shown in Figure 6. This increased density of the field lines produces a magnetic pressure which further displaces the column creating an unstable configuration [41][pg 118]. The condition for the kink instability is derived from ideal MHD and requires that $q > 1$ in the plasma to prevent kink instabilities. As q can be inversely related to the poloidal field, this places limits on the current that can be driven through toroidal devices [56].

2.2 Tokamak Disruptions

A clear aim of tokamak operation is to maximise the power output achievable during operation. One can show from the thermonuclear power in a D-T plasma that the fusion power, P_{fus} depends on

$$P_{fus} \propto p^2 V \propto \beta^2 B^4 V, \tag{2.22}$$

so that fusion power depends on the plasma beta, the magnetic field and the volume of the device [56]. Increasing the size of the device is planned for future tokamaks like ITER, but increasing the size or magnetic field also increases the cost [56]. Another way to improve performance is by increasing the plasma pressure. More generally, experimental scaling laws have been derived which show how performance depends on different engineering or plasma parameters

$$\tau_E \propto I^{0.93} B_\phi^{0.15} P^{-0.69} n_e^{0.41} M^{0.19} R^{1.97} \epsilon^{0.58} \kappa^{0.78} \tag{2.23}$$

where I is the plasma current, P the applied heating power, n_e the line averaged electron density, M is the isotope mass, $\epsilon = a/R$ is the inverse aspect ratio and κ is the plasma elongation [33]. The plasma elongation refers to the shape of the plasma being longer in the vertical direction in a divertor tokamak. This equation highlights how confinement times can be improved, thus improving the fusion triple product given in Equation 1.4.

In practise, the parameter space for reliable tokamak operation is limited by a variety of constraints and the listed parameters cannot be increased indefinitely [45]. Increasing plasma parameters can excite MHD instabilities whose effects can impact performance or in more extreme cases result in a disruption causing termination of the plasma [19]. One such condition is the ideal beta limit which states that the maximum achievable plasma beta, β_M is given by

$$\beta_M = \frac{gI}{aB_\phi} \tag{2.24}$$

where g is the Troyon factor [56]. Plasma betas are normalised by this factor so that

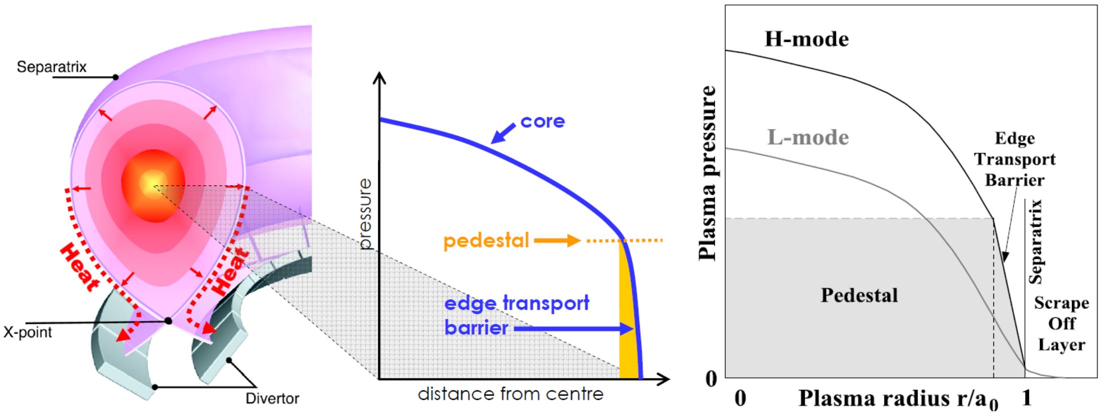
$$\beta_N = \frac{\beta a B_\phi}{I} \tag{2.25}$$

and the stabilised condition is simply $\beta_N < g \simeq 3.5$ from ideal MHD [34]. MHD instabilities impose limits on both the maximum plasma pressure and the maximum plasma current. The current limit can be expressed in terms of the safety factor which in a cylindrical approximation can be written

$$q_a = \frac{2\pi a^2 B_\phi}{\mu_0 R I} \tag{2.26}$$

so that a maximum current limit corresponds to a minimum value of the safety factor [45]. The MHD limit is typically cited as $q_a < 2$ results in the destabilisation of what is known as an external kink mode [56], though the role of q is debated [136]. The presence of MHD instabilities, in particular edge-localised modes, tearing modes and sawteeth are expected to play an important role in the achievable performance of ITER [34].

2.2.1 Edge Localised Modes and Numerical simulations



(a)

(b)

Figure 7: a) Illustration of the magnetic topology of a tokamak with an X-point in the poloidal field with heat flux travelling along the outer most field lines to the divertor, reproduced from [62]. Additionally, the characteristic pressure profile from the core of the plasma to the outer region is also shown. b) Characteristic pressure profiles for L and H-modes against the plasma from the centre to edge. Transition from L to H-mode is characterised by large pressure gradients in a narrow region. Figure reproduced from [93].

Experiments conducted in the ASDEX tokamak at Garching, Germany by [10] showed that above a given threshold in heating power, a high-confinement mode (H-mode) could be achieved. Plasma confinement time in this new H-mode is observed to be a factor of 2 or higher [96]. The high-confinement mode brought an improved transport barrier in a narrow region at the plasma edge, resulting in reduced particle and heat transport across the plasma edge. This transport barrier is typically several centimetres wide, and characterised by steep pressure and density gradients at the plasma edge [52]. Figure 7 shows a comparison between the previous low confinement (L-mode) and the H-mode with larger pressure gradients.

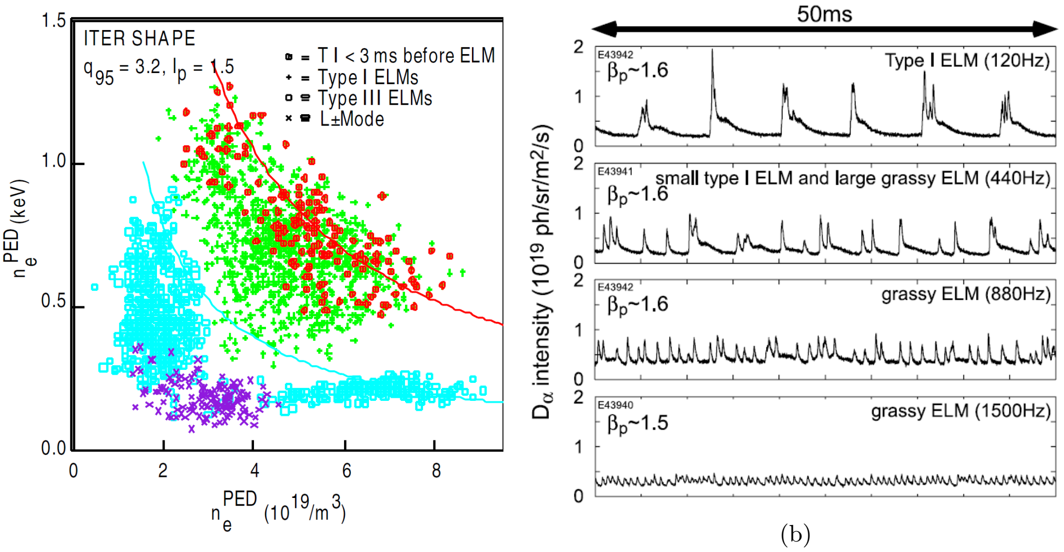
However, while this new H-mode brought improved confinement properties, it was also accompanied by the appearance of instabilities and plasma wall-interactions. These instabilities appeared to have largest amplitude in these high pressure regions and accordingly given the name edge-localised modes (ELMs). Such behaviour was found to be characteristic of X-point divertor tokamaks, shown in Figure 7. In this tokamak configuration, the core of the plasma is characterised by closed magnetic field lines or flux surfaces while the outer region, or scrape off-layer is separated by an open magnetic field-line which connects to the divertor. This configuration allows for particle exhaust from the plasma to the divertors. The Figure also shows the corresponding radial pressure profile in the tokamak, with large pressure gradients towards the plasma edge [62].

ELMs result in particle emissions at the transport barrier, lowering the pressure gradients and temperatures in this region. These ejections result in loss of $10 - 15\%$ of energy in the transport barrier [52]. Following this instability the plasma begins to recover and pressure gradients and temperatures rise again until the process repeats. The time between these repeat crashes then gives the frequency of the instability. ELMs can then be thought of as quasi-periodic disturbances which occur at the plasma edge, on time-scales of the order of μs [28]. The current theoretical understanding of ELMs comes from ideal magnetohydrodynamics [96].

While the loss of confinement is not ideal, of greater concern is the erosion these ejections can cause to the tokamak itself. It is thought that the severity of these ejections scale with tokamak size and will then be detrimental to the operation of large future tokamaks like ITER [93]. Further, these ejections can cause radial transport toward the main walls of the reactor, and result in liberation of impurities from the tokamak walls which cause further radiative heat loss. It is important for future ITER operation to understand not only the causal effects of ELMs, but ways to successfully control them [96]. Thus, it would seem beneficial to have an ELM free regime, with high pressure and temperature gradients resulting in good confinement. However, ELMs can offer some benefits in removing impurities from the plasma and hence there is also interest in regimes with small ELMs and high confinement properties [62].

ELMs are classified into several distinct types, first owing to Doyle and experiments taken at the DIII-D tokamak in 1995 [28]. The most commonly observed ELMs are classed as Type I and Type III. The difference between the two is typically measured by their repetition frequency dependence on supplied heating power. Type I ELMs show an increasing repetition with increasing heating power, while Type III ELMs show a decreasing repetition frequency with increased heating power [96]. In general, it seems that Type I ELMs have better confinement properties, but result in larger expulsions of energy compared to Type III ELMs. Figure 8 shows the emission of D_α light from the JT-60U tokamak. This D_α is visible red light which arises from the interaction of emitted electrons from the plasma with neutral particles [96]. During an ELM, there is an observed peak of emitted D_α in a small period before it returns to zero.

Another way of classifying ELMs comes from observation of the density (n) and temperature (T). Figure 8 shows measurements taken at the DIII-D tokamak at the inner edge of the transport barrier. The red and blue curves show lines of constant pressure. We can see that Type I ELMs are clustered along lines where $T_e n_e \sim$ constant. This line corresponds to a line of pressure at the onset of what is known as the ballooning instability, implying that pressure plays some role in their evolution [62]. Type III ELMs are clustered in two separate regions. The first is found along a line of constant density, and the second along a line of constant pressure. Both clusters occur in regions of pressure well below Type I ELMs [96].



(a)

(b)

Figure 8: a) Plot of temperature vs density for different ELM observations showing a separation of ELM Types in this space, reproduced from [37]. b) the D_α light emission for different ELM Types recorded at the JT-60U tokamak. This shows large Type-I ELMs, and smaller grassy Type ELMs, reproduced from [53].

Type I ELMs result in a much greater expulsion of heat and energy than Type III ELMs, but typically occur at high pressures. It then becomes desirable to find an operational regime where smaller expulsions similar to Type III ELMs occur at pressures close to Type I ELMs. Several ELMs like this seem to exist including; Type I, Grassy ELMs and Type V. However, these types are not as well understood [96].

Another feature of ELMs is an associated filamentary structure, where these filaments are aligned with the magnetic field, and extend radially up to $5 - 10$ cm showing typical toroidal mode numbers of $n \sim 10 - 15$. Visual observation of these filamentary structures were taken at MAST [60], shown in Figure 9, where the plasma surface is viewed through a port [96]. The filaments are aligned with the magnetic field.

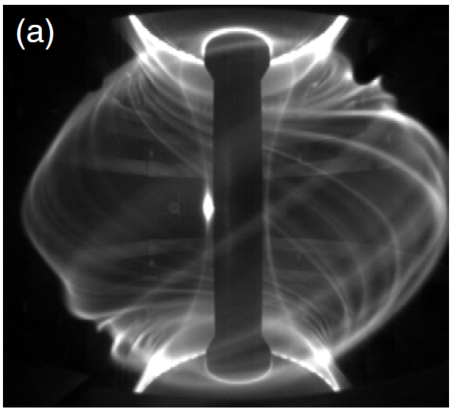
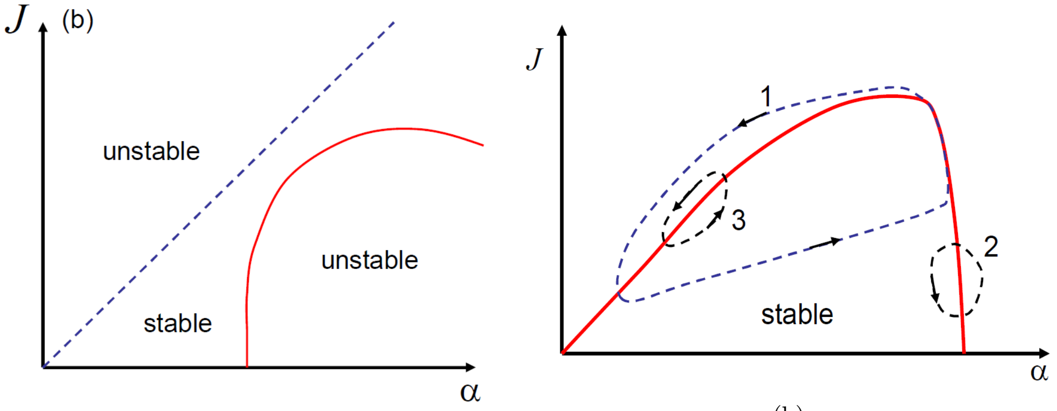


Figure 9: Filamentary observation of ELMs at the MAST tokamak which demonstrates plasma ejections shown in white at the scrape-off layer, reproduced from [60].

The Peeling-Ballooning Model



(a)

(b)

Figure 10: a) Stability diagram for the ballooning (red) and peeling (dashed blue). b) A simplified representation of the peeling-ballooning stability space with current density J and normalised pressure gradient α. The dashed lines show possible trajectories representing different ELM Types. Below the red curve are stable parameter values, and above are unstable. Figures reproduced from [96].

Over the years, there has been increased support that ELMs are the result of the coupling of an instability in pressure gradients called the ballooning instability, and current density at the plasma boundary called the peeling instability [52]. This resulted in the so-called peeling-ballooning model, which combines the effects of these two instabilities. The peeling-ballooning model has mainly been used to describe the onset of Type I ELMs [86].

The ballooning instability has a long wavelength parallel to magnetic field lines, and a short one perpendicular. Pressure gradients drive this instability, but it is stabilised by current density. This can be completely stabilised if the current density is high enough, and is called “second stability access”. Conversely, the peeling instability is destabilised by current density and stabilised by pressure gradients. However, at high pressure gradients, the two instabilities can couple forming the peeling-ballooning instability. It is the coupled instability which is thought to be responsible for Type I ELMs [96]. The left of Figure 10 shows a qualitative representation of these effects on stability. The dashed blue line shows the peeling instability threshold, and the solid red curve represents the threshold for the ballooning instability.

Right of Figure 10 shows a simplified schematic of the parameter stability for the peeling-ballooning model. The discussion and Figure are taken from [96]. We can see that for sufficiently high pressure gradients and low current densities, the model is driven unstable by the ballooning instability. Similarly at higher current density, the peeling mode can drive instability. The dashed curves labelled 1,2 and 3 show possible trajectories in this parameter space representing different possible ELM Types.

Trajectory 1 represents a Type I ELM. While the plasma is stable, the pressure gradients and current density continually increase until the ballooning stability boundary is reached. The current then continues to rise until the combination of high pressure gradient and current density crosses the peeling-ballooning stability boundary. Once this stability limit has been crossed, the instability grows exponentially, causing significant energy transport from convection or conduction [93]. The pressure gradients and currents decrease, removing the drive of the instability, until the trajectory again falls into the stable parameter space where the process repeats. In general it takes longer for the currents to diffuse than the pressure [62].

Other trajectories have been suggested for explaining different ELM Types. For case 2, the pressure gradient is high, but only the ballooning instability is destabilised. The comparatively low current gradient means that the ELM is quickly stabilised without a large loss of pressure. This is intended to represent smaller ELMs. Finally trajectory 3 is close to the peeling stability boundary, and results in a larger loss of pressure. The pressure gradient required for this trajectory is much lower than 1 or 2, and hence could explain Type III ELMs. However, these latter two cases are not as widely accepted [96].

Several different numerical codes exist for studying the linear and nonlinear behaviour of the peeling-ballooning model. For the linear stability, the ELITE code [40] is one example. Linear stability analyses are useful to understand the onset of instability on a plasma. For instance, they may answer when we expect the instability to appear. However, understanding how much energy is lost during the evolution is in general not possible. The linear stability of the peeling-ballooning model has been studied by codes such ELITE, GATO and MISHKA [83]. Other codes have been developed to study the nonlinear development of the peeling-ballooning model such as JOREK [131] and BOUT++ [73, 83].

2.2.2 Sawtooth Instability

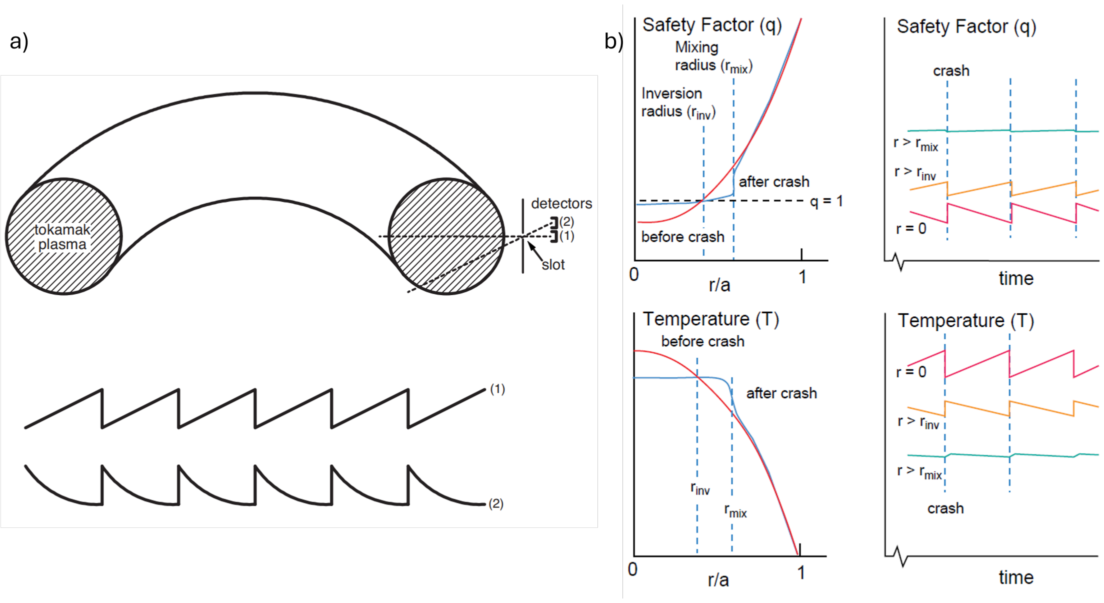


Figure 11: a) Schematic diagram of a sawtooth measurement taken in an idealised tokamak with soft X-ray measurements. Figure reproduced from [79]. b) Evolution of the temperature profile and q in an idealized tokamak. Figure reproduced from [34].

The sawtooth instability is a large scale plasma instability that causes periodic relaxations in the core plasma temperature and density. The instability was first observed in 1974 in the ST tokamak by ref. [4] and has since been observed in every subsequent tokamak. A measure of the plasma temperature through electron cyclotron emission [16][pg 244] shows a rise in the central plasma temperature followed by a sudden crash, shown in Figure 11 b). When measured at the plasma edge, the temperature measurements are inverted meaning a decrease in core temperature results in a corresponding increase in temperature at the plasma edge. The radius at which the temperature remains approximately constant is referred to as the inversion radius, r_{inv} [77].

A sawtooth cycle consists of 3 different phases: a ramp up phase which is associated with an increasing temperature profile in the core, a precursor phase where the plasma becomes unstable to a kink mode instability and finally a fast crash where the temperature drops rapidly. This behaviour can be understood in the following way [41][pg 160]. The cycle begins with the temperature profile being relatively flat and the safety factor $q > 1$ everywhere. As the current is peaked on-axis, the plasma experiences a higher degree of Ohmic heating in the core. The higher temperature leads to an increase in conductivity of the plasma in the core and this in turn leads to higher currents. An increase in currents generates more poloidal field which results in the safety factor $q < 1$ allowing a $q = 1$ flux surface to form. This surface is then susceptible to an $m = 1$, $n = 1$ internal kink instability whose nonlinear evolution leads to a crash in temperature at the core of the plasma [136]. Following the sawtooth crash, mixing of the plasma occurs within the radius r_{mix} and the temperature profiles are flattened. The cycle then repeats itself when $q(0) > 1$ is restored. While this is a useful method to understand the sawtooth it should be noted that this is not a complete description [77]. A comprehensive discussion of different models is given by [158].

As the sawtooth instability involves large-scale movement of the plasma, it can be observed experimentally through many different diagnostics including: soft X-rays, electron cyclotron emission and density measurements. Due to plasma toroidal plasma rotation, there can be a precursor phase before a crash shown in Figure 12. As the kink mode grows in amplitude, this appears as a growing oscillation in the laboratory frame of reference. It should be noted through that precursorless sawteeth oscillations have also been observed [14].

Many models exist to explain sawteeth including finite reconnection models, two-fluid models and kinetic models ([22][pg. 246] and references therein), with the exact mechanism still remaining disputed many years after the first observations of sawteeth [136]. The first model termed the Kadomtsev model [5] attempts to explain the process in terms of helical magnetic flux formed by the $q = 1$ surface [16][pg176]. The model suggests that magnetic reconnection will occur for the $q = 1$ surface until this new island is completely annihilated and a stable equilibrium is formed again with $q > 1$. However the model does not explain the quiescent ramp phase observed in the sawtooth and also does not accurately capture the timescales observed in the sawtooth instability [30]. Further, the role of magnetic reconnection in causing the sawtooth crash is also debated [136] and experimentally it has been observed that $q(0) < 1$ for the duration of a sawtooth crash [22][pg. 247] which has led to modifications of this theory. The work of ref. [136] proposes a novel method for explanation for the sawtooth instability, but involves complicated long-time MHD simulations and so we merely note the difficulties involved in formulating a consistent theory.

Sawteeth with an inversion radius less than 40% of the tokamak minor radius and temperature drops in the order of a fraction of a keV are tolerable in operation [77]. Indeed the instability can be beneficial for removing impurities in the core which result in energy loss through radiation. While sawteeth themselves do not typically cause a termination of the plasma, sawteeth with a long period can couple to more deleterious instabilities such as edge-localised modes and neo-classical tearing modes which can result in termination of the plasma [77]. Further sawteeth with sufficiently large amplitude can cause loss of energetic α particles from the core and thus limit the tokamak performance. In ITER the presence of fusion α particles is expected to lead to long sawtooth periods [27]. ITER is expected to operate in a so-called "sawtoothing ELMy H-mode" below the Troyon limit. However, there is concern that the coupling of the sawtooth to other instabilities will significantly lower this stability boundary and thus the achievable β_N [34]. As such there is much interest in the control of this instability, particularly through the stability of the internal kink mode, both stabilising and intentionally de-stabilising it.

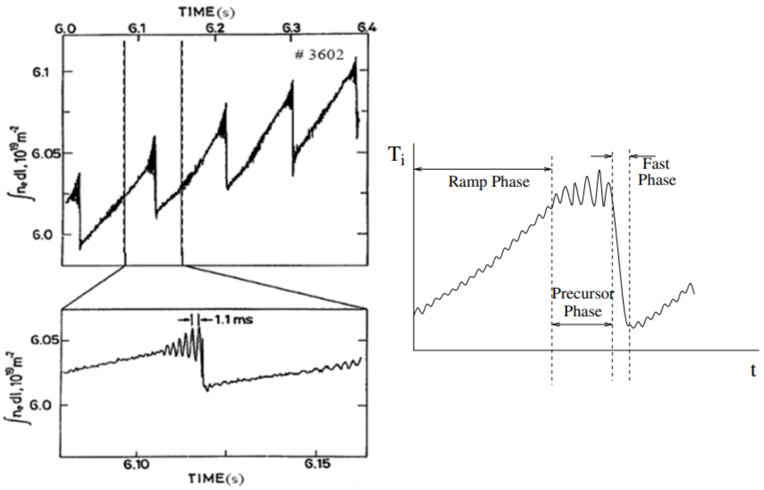


Figure 12: The left of the figure shows the line-integrated density of a sawtooth oscillation observed in JET and the right of the figure highlights a typical sawtooth cycle. The sawtooth cycle is characterised by a slow ramp up phase followed by a sudden crash in temperature which can sometimes be preceded by a precursor phase. Figure reproduced from [77]

2.3 Tokamak diagnostics

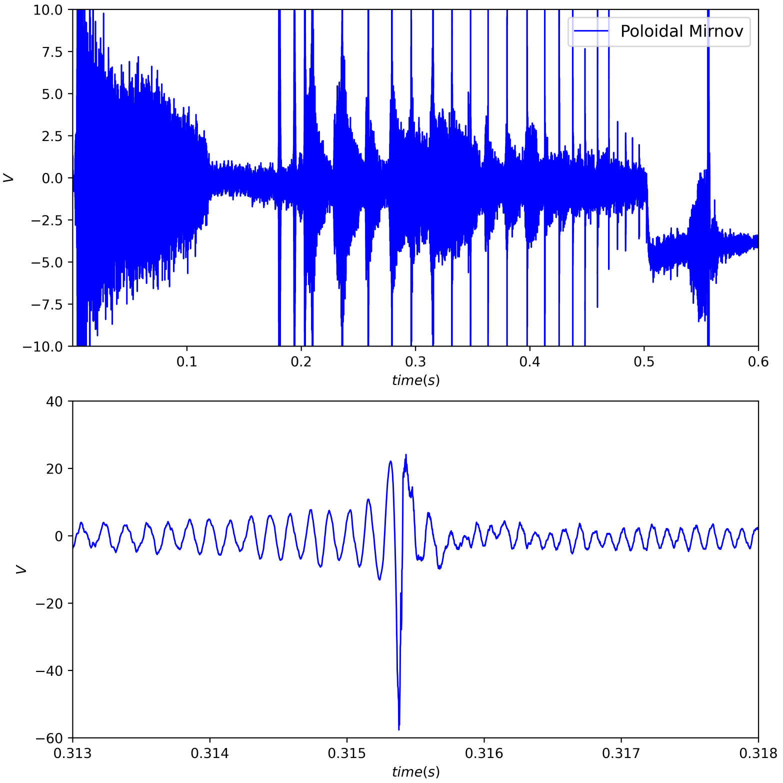


Figure 13: Mirnov measurement made on the MAST tokamak during a sawtoothing event (shot no. 29880). The top figure shows the Mirnov signal for several sawtooth crashes in core temperature and the bottom shows the same signal over a shorter window around a single crash.

Tokamaks contain a large suite of diagnostics which measure an impressive range of different plasma parameters such as density, temperature and magnetic fields. Here we briefly mention some of the diagnostics relevant this dissertation which measure the local magnetic field or the core temperature. These measurements are related to the sawtooth instability described in the previous section. The diagnostics are mentioned to underline both their qualitative features during discharges and the expectation of noise and sampling rates which will be relevant in later sections of this dissertation.

A typical MAST/MAST-U magnetic diagnostics time trace is shown in Figure 13 underpinning a number of challenges. This Mirnov signal shows a sawtoothing event between $t \approx 0.27s$ and $t \approx 0.5s$, and it can see that spikes in the signal are observed on extremely fast timescales which correspond to central temperature crashes (the corresponding soft X-ray is shown in Figure 16 later). As such suitable diagnostics must be selected which have sufficient temporal resolution of these timescales. Further, the data is noisy, and it is far from clear that the noise is simply Gaussian in nature.

2.3.1 Magnetic diagnostics

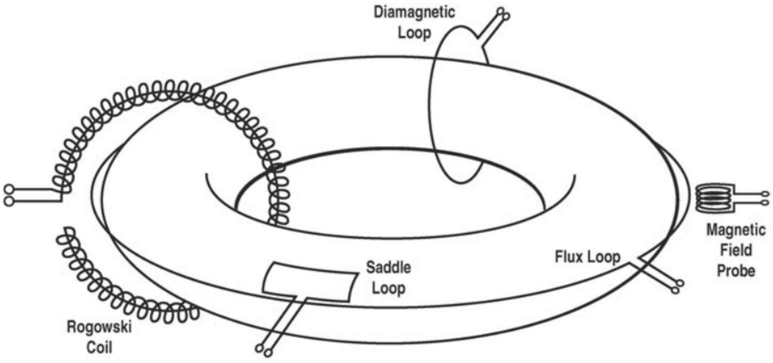


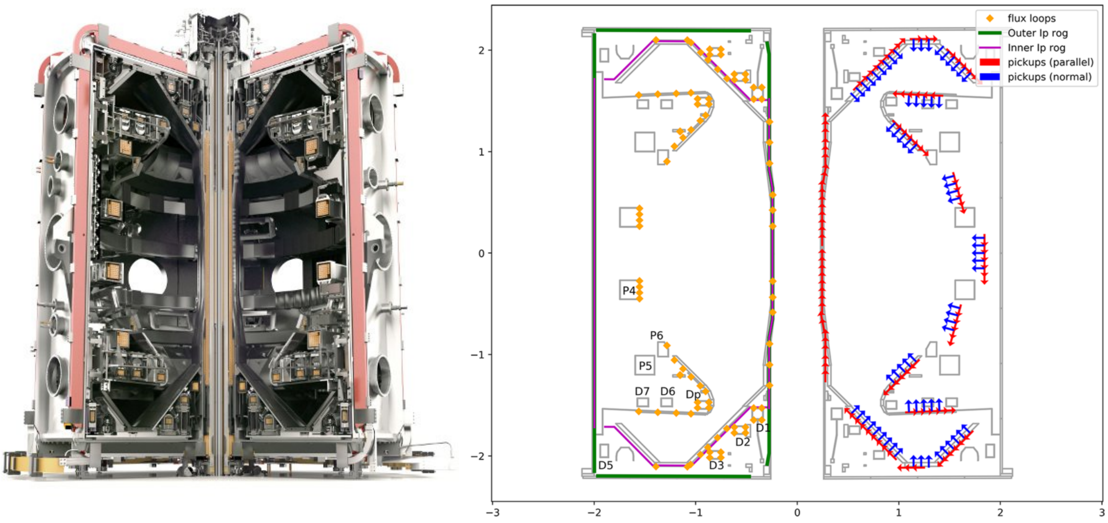
Figure 14: Schematic of a tokamak with some examples of the magnetic diagnostics: poloidal flux loop, magnetic field probe, saddle loop, diamagnetic loop, and Rogowski coils. Figure taken from [63].

Magnetic diagnostics are used for measuring currents, magnetic fields and magnetic fluxes in a tokamak. A general schematic overview of some common diagnostics is shown in Figure 14. Measurements of the magnetic field are important in determining, for example, the equilibrium field of the plasma or observing oscillations in the plasma and diagnosing MHD instabilities. In our case, measurements of the magnetic field are of interest and tokamaks typically take many local spatial measurements of both the toroidal and poloidal magnetic fields.

A Mirnov coil or sometimes called a pickup coil consists of a conductive wire wrapped into a solenoidal coil measuring the poloidal or toroidal magnetic fields. The changing magnetic field of the plasma induces a current in the solenoid which gives a measure of the local rate of change of the magnetic field. As these coils measure high frequency plasma oscillations, they are located on the internal vessel wall, otherwise conducting materials can attenuate these high frequency signals [63]. This means that diagnostics are typically shielded with graphite to protect them from particle flux.

Arrays of these coils are located throughout different toroidal and poloidal locations on the inner vessel wall. MHD modes can appear as oscillations in the magnetic field, and the mode numbers of the instabilities can be determined from phase differences between different Mirnov locations. MHD instabilities can rotate toroidally in the laboratory frame of reference which allows them to be detected due to the rate of change of magnetic field. However, even modes which are slowly propagating or stationary with respect to the plasma still show in magnetic measurements due to the rotation of the plasma with respect to the laboratory frame of reference. This plasma rotation largely results from the momentum transfer from neutral-beam injection commonly used to heat the plasma.

In MAST-U, the median error of pickup coils is cited around 6.3% (though many are higher). This ratio is calculated over all probes from as the ratio of an expected measurement to an observed measurement in a calibration run and measure at rates of 200kHz [166]. In reality, many other factors can contribute to error such as quantisation error when recording numeric values. This tokamak features up to 354 pick-up coil measurements throughout the device, though in most shots not all of them are used. An example of the pickup coil locations is shown in Figure 15 b).



(a) (b)

Figure 15: a) The MAST-U tokamak cross-section and b) example locations of magnetic diagnostics through a poloidal cross-section (reproduced from [166]).

2.3.2 X-ray measurements

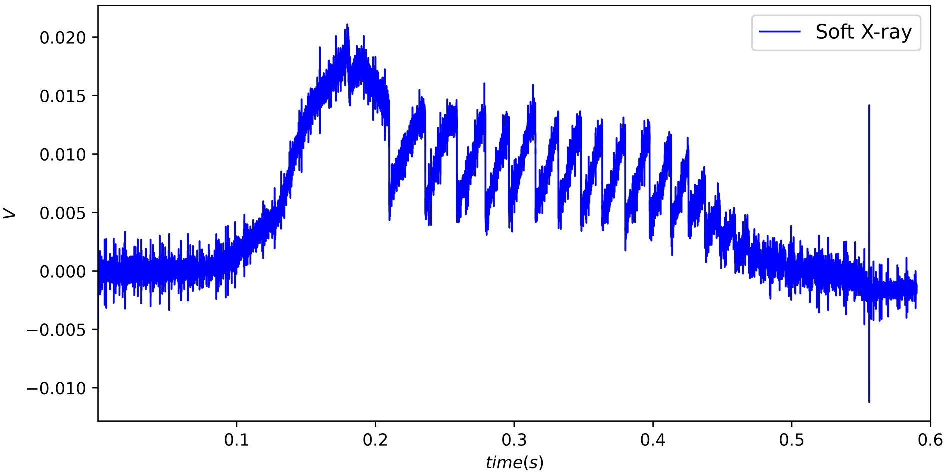


Figure 16: Soft X-ray measurement made on the MAST tokamak (shot no. 29880) with a measurement taken over an entire shot lasting 0.6s. At approximately $t = 0.2s$, a sawtooth instability takes place and can be seen by the sawtooth-like teeth shown in the measurement.

A hot plasma consisting of electrons and hydrogen ions emits electromagnetic radiation due to collisions. This radiation results from braking or bremsstrahlung radiation due to electron-ion collisions in the plasma [16]. The emission spectrum is continuous for a Maxwellian distribution of electrons and can be written

$$\epsilon \propto n_e^2 Z_{eff} \sqrt{\frac{1}{T_e}} \exp(-hv/T_e) \tag{2.27}$$

where ϵ is the emissivity, hv is the photon energy and Z_{eff} is the effective charge of the plasma [64]. This shows that emissivity is related to the plasma electron temperature.

The plasma also contains a number of impurity ions resulting from plasma-component interactions which also radiate power from the plasma. In this case, line emission can be observed from bound electron transitions where electrons are excited to higher bound energy levels through collisions. Transitions to lower energies then causes radiative emission of distinct frequencies related to the energy level of the transitions.

Measurements of both continuum radiation and line radiation can be made using spectroscopic techniques. These measurements are made by spectroscopic cameras along a line directed into the plasma, so measurements correspond to the line integration of the emissivity in the observed volume. We are primarily interested in soft x-rays which are predominantly due to Bremsstrahlung radiation and so measure the plasma temperature in the core. This makes them suited for measuring the evolution of the equilibrium and sawtooth instability.

In MAST-U there are two soft X-ray cameras with 14 lines of sight each [174]. These cameras measure the soft X-ray emission in different locations throughout the plasma. There is no clear description given in the MAST-U diagnostics handbook of the exact details of these daignostics or their sampling rates. However, they are typically sampled at high frequency and, on occasion, we have observed that they are sampled as frequently as the Mirnov coils at 200Hz. We typically refer to the soft X-ray measurements for the temperature of the plasma as the sampling rate is generally much higher than other diagnostics. An example measurement of the soft X-rays for a short time is shown in Figure 16. More information is available on the soft X-ray diagnostics in MAST which consist of cameras measuring upper horizontal, lower horizontal and tangential plasma. For this report, the tangential camera is the relevant camera which measures between the core and the plasma edge [151]. These cameras filter line emission below a certain energy corresponding to impurity radiation. The detected signal is then primarily due to bremsstrahlung radiation from the plasma core [16].

2.4 A symmetry based view of gross Tokamak behaviour

We now introduce the derivations of the bifurcation theory models discussed in ref. [72] with some associated extensions. These models will come to be referred to regularly and remain an important reference for later sections of the dissertation.

2.4.1 Taylor-Couette Flow

Tokamak instabilities are complicated and nonlinear in nature, and there is naturally interest in developing simplified models of their behaviour [71]. Insight can be gained by drawing parallels between the tokamak and related fluids experiments, specifically Taylor-Couette flow [72]. Before discussing the derivation of these nonlinear bifurcation models, we first describe Taylor-Couette flow, an experiment studied by [1]. The description given here follows closely those given by [85] and [23].

Taylor-Couette flow consists of a fluid with viscosity ν bound between two concentric cylinders with radii r_0 and r_1 where either one or both of the cylindrical surfaces can be made to rotate at angular velocity Ω_0 and Ω_1. As the speed of rotation is increased, the flow undergoes a distinct change or bifurcation with the appearance of rolls named Taylor vortices. These rolls are in counter-rotating pairs and appear to be stacked on top of one another remaining the same around the annulus. As the speed further increases these rolls become wavy around the circumference of the device, forming standing waves. A further increase in speed causes the appearance of modulated wavy vortices which are no longer steady state as the waves rotates around the device. Finally there is a transition from modulated wavy vortices to turbulent wavy vortices which ends in a transition to featureless turbulence. Some of the main sequences are shown in Figure 17.

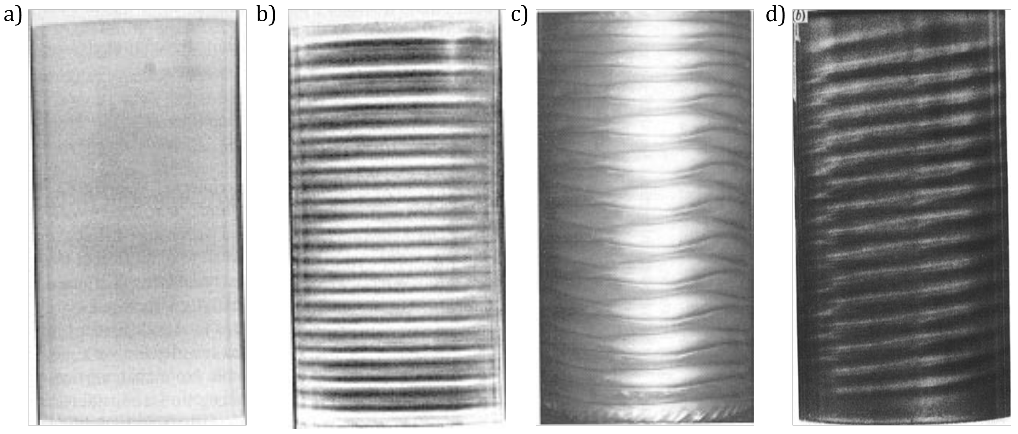


Figure 17: Taylour-Couette Experiment showing different behaviours. a) Couette flow or shear flow, b) Taylor vortices, c) wavy vortices, d) spiral vortices. Adapted from [36].

The exact transitions to these different states can depend on many different factors. For instance, Figure 17 d) shows a case where the inner and outer cylinders rotate at different speeds. This allows a completely new type of spiral pattern to form which can also become modulated in the same way as Taylor vortices. Further, the way in which the final cylinder speed is reached can also impact the resulting pattern formation. The formation of these patterns can be understood in terms of the symmetries apparent in the Taylor-Couette experiment and the progressive loss of these symmetries. We first describe the symmetries of the Taylor-Couette experiment.

If we consider the experimental apparatus as two finite concentric cylinders we can write down two relevant symmetries. The first is an azimuthal symmetry where rotating the device by some angle leaves everything unchanged, this is denoted by the group of special orthogonal rotation in 2D, $SO(2)$. The next symmetry is a reflection symmetry in the axial direction denoted by the parity group $Z(2)$. That is, if the experiment were reflected about the bottom plane, all rotations would remain in the same direction. However, we can include a further symmetry if we consider the experiment to consist of an infinite cylinder in the axial direction represented by periodic boundary conditions. This is given by a further $SO(2)$ group if the cylinder is considered to be a high aspect ratio torus such that the two ends are joined to one another. The product of the $Z(2)$ group and the $SO(2)$ groups in the axial direction forms the larger group $O(2)$ having all transformations with determinant ± 1 which preserve angles and distances.

The resulting pattern formation can now be understood in terms of a progressive breaking of these symmetries, either entirely or with only a subgroup of the original group remaining. Initially, the first shear state obeys all the outlined symmetries, invariant in the azimuthal and axial directions and under a parity reflection. The transition from shear flow to Taylor-vortices results in the loss of the $O(2)$ symmetry in the axial direction. The one symmetry retained in this direction is a smaller subgroup of the $O(2)$ group given by rotations in the axial direction spanning the height of a vortex pair, but we retain all other symmetries.

The appearance of wavy vortices results in the further loss of the azimuthal $SO(2)$ symmetry and the parity symmetry $Z(2)$ while retaining the subgroup of the Taylor-vortices. However, while a reflection causes the appearance of the wave to change, a combination of reflection and rotation can return the Taylor-Couette flow to its original state. We can also see from Figure 17 that wavy flow retains some periodicity in the azimuthal direction and so rotations by a fixed degree still leave the system invariant. Another symmetry is given by the fact that wavy vortex flow appears to rotate rigidly. This signifies a time-periodic symmetry which after one period the system returns to its original state. The final symmetry is given by a combination of both spatial and temporal symmetries where if the system is observed in a frame which rotates with wavy vortices, it again appears steady-state. The transition to modulated wavy vortices results in the loss of these spatio-temporal symmetries and they become more complicated in their symmetries.

So we can think of the allowable patterns as different solutions to the Navier-Stokes equations which obey different symmetries. The modes which form stably depend on the rotation rates of the inner and outer cylinders, with Taylor-vortices forming if the outer cylinder is counterrated at a speed less than a given critical value. Thus symmetry tells us something fundamental about the behaviour the solutions must obey, regardless of the model. While the physics may downselect some of the permissible modes, symmetry plays a key role in the solutions which can appear.

2.4.2 A symmetry based view of gross Tokamak behaviour

We now discuss the application of these ideas to the nonlinear modelling of tokamak behaviour developed by [71] and [72]. Modeling of tokamak behaviour is challenging for a number of reasons. We list a few of these given by [71]: tokamaks are by nature nonlinear from the balance of fluid pressure to the Lorentz force which is quadratic, tokamaks can have sophisticated control mechanisms which shape the plasma and experimental measurements are challenging in fusion conditions. Despite these set-backs, bifurcation theory could aid in the topological nonlinear study of these complicated instabilities as tokamaks often operate close to instability boundaries.

The geometry of the Taylor-Couette experiment shares many of the symmetries of a tokamak with circular cross-section and large aspect ratio. As described in the previous section, the symmetry of the group of the Taylor-Couette experiment can be written

$$G_s(T-C) = SO(2) \times (SO(2) \ltimes Z(2)) = SO(2) \times O(2)$$

where $\ltimes$ is the semi-direct product. In analogy to the tokamak, a similar group of symmetries can be written

$$G_s(Tok) = (SO(2) \times SO(2)) \ltimes Z(2),$$

is the symmetry group of the periodic cylinder model of a tokamak. There are two $SO(2)$ groups, the first corresponding to rotation in the poloidal direction and the second to rotations in the toroidal direction. However, in general most tokamak designs have non-circular cross-sections with the feature of divertors which create X-points in the plasma so we will tend to ignore this symmetry. The final group can be understood by considering a mirror reflection of the periodic cylinder at one end. The applied magnetic fields are reversed but the tokamak is invariant to reversal of the current in the toroidal direction implying the dynamics remain the same. A parity reflection of this type corresponds to a co-ordinate transform $(\theta, \phi) \to (-\theta, -\phi)$. The final symmetry we consider is a time-invariance given by energy conservation from ideal MHD.

It is apparent that the tokamak and Taylor-Couette experiment share similar symmetry groups. Further, Taylor-Couette flow is a four parameter system characterised by the inner and outer radii with driving energy coming from the rotation of the outer and inner cylinders. An Ohmically heated tokamak can also be thought of as a four parameter system where the two radii correspond to minor and major radii of the torus and the driving energy is given by the imposed magnetic field and the Ohmic heating from the rate of change of the magnetic field (or the induced current in the plasma). Possible analogies of the different behaviour in Taylor-Couette flow have been drawn by [72] and are summarised in Table 1. In particular, between the Taylor-vortices and the sawtooth both are non-traveling waves with rolls being prevalent in the TC experiment and the sawtooth also being ubiquitous [77].

Taylor-Couette	Tokamak
Steady sheared flow	MHD equilibrium
Rolls (Taylor vortices)	Sawtooth oscillation
Rotating wave	Mirnov oscillations
Modulated rotating waves	Complex Mirnov signal

Table 1: Analogies of flow phenomena in the Taylor-Couette experiment to tokamak observations and instabilities.

The analogies between the symmetry groups of the Taylor-Couette experiment and the tokamak lead us to consider the application of equivariant bifurcation theory. This is bifurcation theory with a downselection of terms which do not obey particular symmetries. Bifurcation theory is also often used to describe instabilities close to the instability threshold, for example in weakly nonlinear theory. As already described, tokamaks often operate close to instability boundaries as this provides optimal performance.

We begin by considering a general modal solution consisting of a superposition of complex wave-like solutions

$$y(\phi,t) = a(t)\exp(in\phi) + \bar{a}(t)\exp(-in\phi) \tag{2.28}$$

where a is the complex amplitude of the solution, $\bar{a}$ the complex conjugate and ϕ the axis of symmetry. Given a general solution, we must now ask what form an ODE can take which will obey the outlined symmetry constraint.

If the dynamics are well described by ideal MHD, we will have energy conservation. In this case we start from an explicitly time-independent Lagrangian $\mathcal{L}(y,\dot{y}) = \mathcal{L}(a,\dot{a})$ which satisfies only rotational invariance

$$2\mathcal{L} = \mathcal{T} - \mathcal{V} = |\dot{a}|^2 + \mu|a|^2 + \sigma|a|^4, \tag{2.29}$$

where μ and σ are real valued parameters. $\mathcal{T}$ can be considered as the kinetic energy of the system and $\mathcal{V}$ is the potential in the energy-preserving case. Note that it is entirely possible to expand the potential to higher order, as in [71] the potential is taken to $6th$ order. We express the complex amplitude a in terms of its real amplitude r and phase ξ so we can write $a(t) = r\exp(i\xi)$ and the Lagrangian becomes

$$2\mathcal{L} = \dot{r}^2 + r^2\dot{\xi}^2 + \mu r^2 + \sigma r^4. \tag{2.30}$$

The Euler-Lagrange equations give

$$\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{r}}\right) - \frac{\partial\mathcal{L}}{\partial r} = \ddot{r} - r\dot{\xi}^2 - \mu r - 2\sigma r^3 = 0, \tag{2.31}$$

$$\frac{d}{dt}\left(\frac{\partial\mathcal{L}}{\partial\dot{\xi}}\right) - \frac{\partial\mathcal{L}}{\partial\xi} = \frac{d}{dt}(r^2\dot{\xi}) = 0. \tag{2.32}$$

The second equation yields $\dot{\xi} = C/r^2$ corresponding to conservation of momentum consistent with rotational symmetry. However, to be consistent with the parity symmetry we must have $C = 0$ and are left only with the first of the two equations

$$\ddot{r} = \mu r + 2\sigma r^3. \tag{2.33}$$

As $\xi = \xi_0$ for an arbitrary constant phase ξ_0, we can set $\xi_0 = 0$ and we can write

$$\ddot{a} = \mu a + 2\sigma a^3, \tag{2.34}$$

as $r = a$. This behaviour can be considered as a particle trapped in a potential well. To understand the general behaviour, we consider the onset of instability when a is small. In this case the solutions grow exponentially due to the linear term until a^3 is sufficiently large causing the system to return to small amplitude of a thus exhibits bursty behaviour. The potential well for this model is shown in Figure 18 where the particle can be trapped in the left or right potential wells.

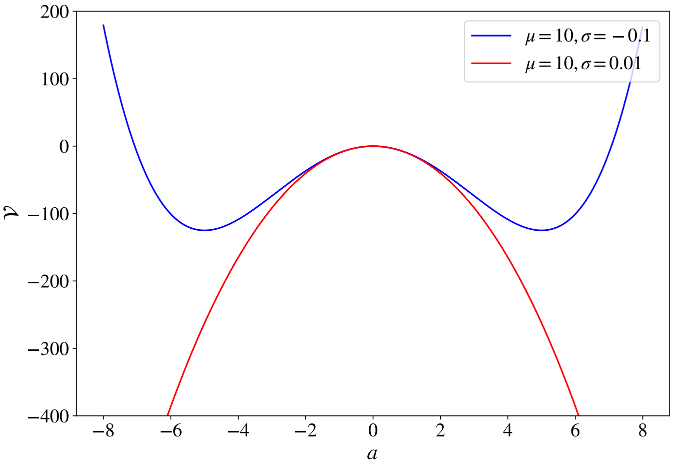


Figure 18: Plot of the ANAC potential well when it is stable with $\sigma < 0$ and unstable with $\sigma > 0$. The behaviour of the trajectories can be understood as the a a particle trapped in a well.

Finally we consider the interaction of Equation 2.34 with an unstable equilibrium mode which is represented by the normal form of a fold bifurcation. To satisfy rotational invariance, interaction of the equilibrium mode with the non-axisymmetric mode will thus be through a term proportional to a^2. We can then write the equilibrium mode as

$$\dot{b} = \alpha - \beta b^2 - a^2. \tag{2.35}$$

The fold bifurcation without the dependence on a^2 gives a system with two fixed points at $b = \pm\sqrt{\alpha/\beta}$ where $-\sqrt{\alpha/\beta}$ is unstable and $\sqrt{\alpha/\beta}$ is stable. Another choice of the coupling term can be made

$$\dot{b} = \alpha - \beta b^2 - (\delta_r b + 1)a^2 \tag{2.36}$$

which models the impact at different major radii in the Tokamak depending on the real valued parameter δ_r [84]. Evolution of the equilibrium mode shows behaviour consistent with sawtooth oscillations, with a slow rise and fast crash which will be discussed in the following section in Figure 25. The coupling of the equilibrium to the oscillator in a is termed the axisymmetric non-axisymmetric coupled model (ANAC).

There are two final modifications to Equation 2.34 which can be considered as extensions to the ANAC model. These terms are included also with a view to fitting experimental signals granting in general a more flexible model [179]. The first is a coupling between the equilibrium mode and a which destroys and recreates the potential well [87]. The system is modified to

$$\ddot{a} = \mu a + 2\sigma r^3 + \mu_2 b a^3. \tag{2.37}$$

The chosen term here satisfies the aforementioned constraints, though its exact form is arbitrarily selected. We now consider the case where $\sigma = 0$ and we consider the case with $\mu < 0$ and $\mu_2 > 0$. The potential well in this case is shown in Figure 19 where for the sake of illustration we treat b as a constant parameter at a given time. At small values of b the potential well is dominated by the quadratic term with coefficient μ for a sufficiently small amplitude of a. The behaviour is therefore a particle trapped in a potential well centred around the origin. As b grows larger, the walls of the potential well are reduced by the term in μ_2 until eventually the particle escapes the potential well causing a sudden growth in a. This sudden growth in a causes a corresponding decrease in b from Equation 2.36. While both equations can produce sawtoothing behaviour, the inclusion of a back-coupling term now means that oscillations on a can be on a faster time-scale than in b.

The final term we include is a high-order nonlinear dissipative term again chosen to the obey rotational invariance

$$\ddot{a} = \mu a + 2\sigma a^3 + \mu_2 b a^3 - \mu_6 a^6 \dot{a}. \tag{2.38}$$

This term represent non-ideal effects though the order of the nonlinearity is chosen such that it is negligible for $|a| < 1$. This term is important with the new back-coupling term as it prevents solutions becoming unstable. The expanded model with the additional terms in μ_2 and μ_6 is termed the axisymmetric non-axisymmetric coupled extra terms (ANAET) model. When applying to experimental tokamak data, it may also be important to include small symmetry-breaking terms to account for control systems [72] or small random terms to model turbulence [17]. In the following sections some parameters will change in notation to stay in line with other published work, but the resulting models are the same as described here.

As a final note, it is important to consider that the models written down here describe general instabilities within a tokamak under the outlined symmetry constraints. While we aim to relate the behaviour to experimental observations, the mode amplitudes can describe any type of instability behaviour observed. Preliminary fitting of these models to experimental time series for sawteeth instabilities (electron temperature measurements) and ELM instabilties (magnetic pick up coils) was performed by [84]. In these case the time series could qualitatively be fit by changing the parameters in the equations manually due to varying oscillation frequencies observed experimentally. This is important, as it suggests that as well as unknown parameters, these may also be non-constant. The model has also been used in synthetic evaluation of RNNs when fitting to experimental ELM measurements [89]. [115] also validated existing EnKF code and performed fitting of the ANAC model to segments of a Mirnov measurement which showed oscillations in the magnetic field. These oscillations were fit to partial observations of the ANAC model resulting in a simple-harmonic oscillator. Alternative approaches using particle filters have been used with the ANAET model, inferring unseen state variables [162], but purely using the ANAET model as a surrogate.

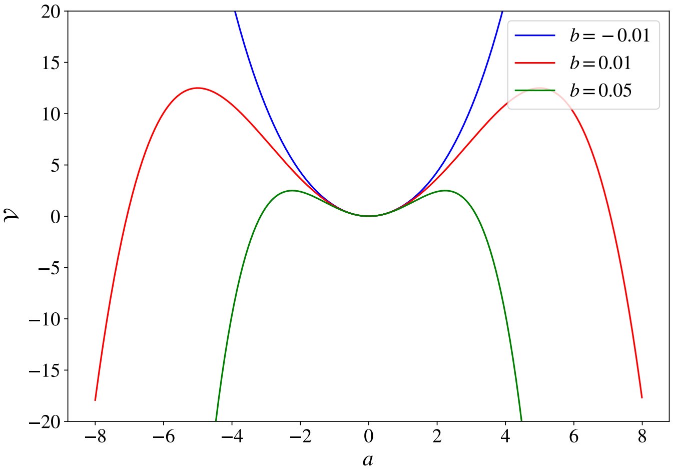


Figure 19: Lowering of the ANAET potential well with varying values of b. The system starts with the blue potential well and as b grows larger, the walls of the potential well lower shown in red then green. This causes the particle to escape over the walls of the potential well.

2.4.3 A qualitative comparison to experimental sawtoothing signals

We now compare qualitatively integrations from the ANAC and ANAET models to some experimental signals to motivate research within this dissertation. We will primarily discuss measurements of the sawtooth instability which is typically observed through the electron temperature, density or soft X-rays. For the sake of data-assimilation against the ANAC or ANAET models, we are also interested in cases where measurements of magnetic activity, measured by magnetic pick up coils, may also be present. This section is a purely qualitative comparison to experiment, but a stricter comparison is described by [17]. In this paper, a bifurcation model is written to describe Mirnov oscillations which are modulated by a slower sawtooth period. Comparisons are then made to the bifurcation model and experimental Mirnov signals.

A report from the JET tokamak in 1992 suggests that "each sawtooth is accompanied by a marked poloidal magnetic field perturbation at the wall" behaviour which is also noted in other Tokamaks of aspect ratio similar to JET $R/a = 2.5$ ([20] and references therein). This perturbation is termed a gong perturbation and was noted to be present in almost all of the sawtooth observations. Figure 20 shows a measurement of the poloidal magnetic field and central temperature taken from the JET tokamak where a correlated spiking in the poloidal magnetic field is observed alongside the sawtooth crash. Figure 21 also shows a separate shot from the JET tokamak showing the poloidal magnetic field and soft X-ray measurements preceding a major disruption in the JET tokamak where again large oscillations in the poloidal field are observed on the crashing frequency of the sawtooth. Qualitative comparisons of the equilibrium mode from the ANAET model in Figure 26 to central temperature measurements show that the equilibrium mode grows in time followed by a rapid crash in a similar fashion to these temperature measurements. The notion of a fast temperature crash is important to experimental observations and not explained by all existing models [14]. Additive noise in the state variable integration of the ANAET model also causes these crashes to occur at different intervals which is also seen in Figure 20.

Of particular note is an observation from the Tokoloshe tokamak which was based in South Africa. Figure 23 shows the response of the $m = 2$ poloidal field to the sawtooth instability. The frequency response shows an increase in frequency of oscillation of the $m = 2$ mode directly after a sawtooth crash with a corresponding increase in amplitude at the crash. Comparison to a noisy integration of the ANAET model in Figure 26 highlights the similarities between the sawtooth instability and experimental results. In particular, at the time of a crash in the ANAET model, the particle escapes the confining potential well causing a sudden crash in b and large amplitude spike in $\dot{a}$. Following this, the frequency of $\dot{a}$ is also increased as the term in μ_2 modifies the shape of the potential well at larger amplitudes [179].

Finally we also plot the ANAC model at varying values of δ_1 in Figure 25. As δ_1 is varied the sawtoothing-like behaviour inverts replicating behaviour seen in the tokamak shown in Figure 22 where the sawtooth inverts at a given radial location in the plasma. As δ_1 is increased to $\delta_1 = -10$ the behaviour becomes ELM like and resembles observed soft X-ray emissions taken near the divertor as shown in Figure 24. We do note an important point about the sawtoothing behaviour in the ANAC model, which is that the sawtoothing period is modulated by a^2 and is on the same period as a. This is distinct from the sawtoothing given by the ANAET model shown in Figure 26, where the slow sawtoothing period is determined by the particle escaping the potential well when b becomes positive. As shown previously in Figures 13 and 16, the oscillations in the Mirnov measurements are on a faster time-scale that the sawtoothing period and so attempts to fit both observations simultaneous would likely imply the need for a back-coupling term. There are, in fact, a number of beneficial features to the ANAET model when fitting generic Mirnov and soft X-ray signals:

1. The number of fast oscillations in a per slow modulation of b can be freely controlled essentially by μ and how quickly b grows towards the fixed points. Specific details are given in [179].
2. Stochastic simulations of the ANAET model generate aperiodic crashing as small perturbations can "push" the particle out of the well earlier / later than in the purely deterministic case. In many cases the sawtoothing is not strictly periodic [136]. In other cases we must vary the parameters to produce aperiodic crashing.
3. High-order diffusion helps solutions remain bounded, returning the particle rapidly to the stable potential well.

With the derivations given in §2.4.2 and comparisons to experiment we have observed that it is possible to derive simplified models which capture key features of the sawtooth instability. These observations motivate fitting equivariant-bifurcation models to experimental time-series of the magnetic field and also soft X-rays. If fits of these surrogate models can be made to experimental signals, it can help reveal the role of symmetry for tokamak instabilities such as the sawtooth. We are then principally interested in cases of fitting $\dot{a}$ to experimental signals or both $\dot{a}$ and b. The experimental observations also underpin a further challenge. These observations are often noisy, non-stationary and poorly sampled. In the remainder of this dissertation we shall discuss approaches for deriving models of this form or fitting these models to data.

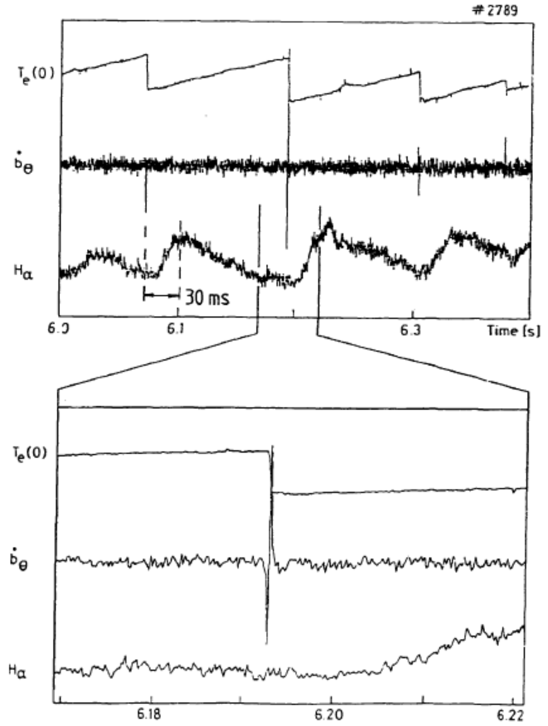


Figure 20: Spikes in the poloidal magnetic field which are correlated with the crashing frequency of the sawtooth on JET. The maximum of the H_α pulse occurs later indicating heat reaching the plasma edge at a later time. Figure taken from [18].

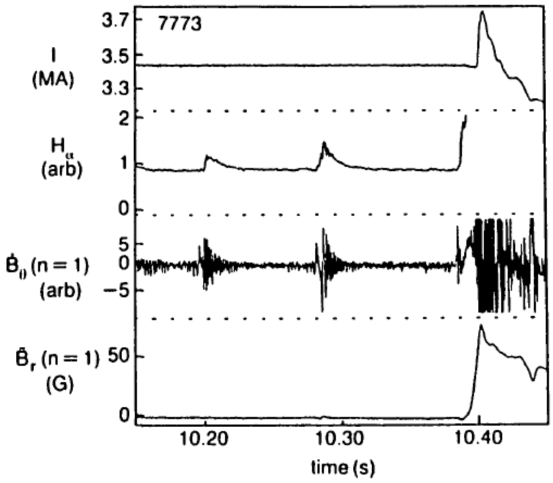


Figure 21: In order: the time behaviour of the plasma current, soft X-ray measurement, at the edge, poloidal field, radial field. Sawteeth are present at $t = 10.2s$ and $t = 10.3$ before the shot is terminated by a major disruption. Figure taken from [19].

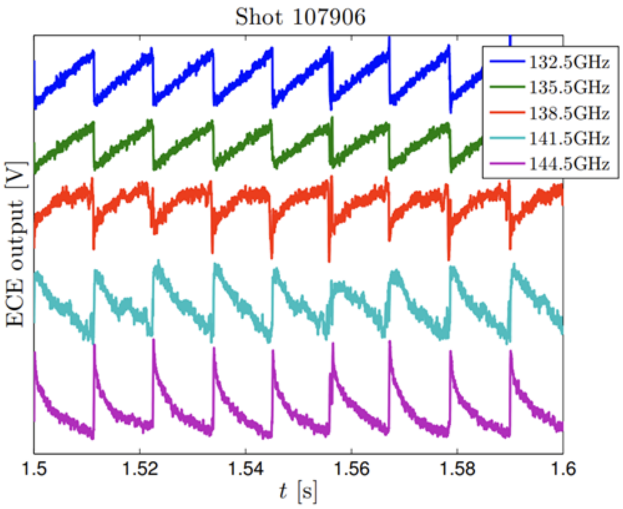


Fig. 8: ECE measurements of TEXTOR shot 107906

Figure 22: ECE measurements of TEXTOR shot 107906, where measurements on different frequncy channels represent different radial locations in the plasma. Figure taken from [76].

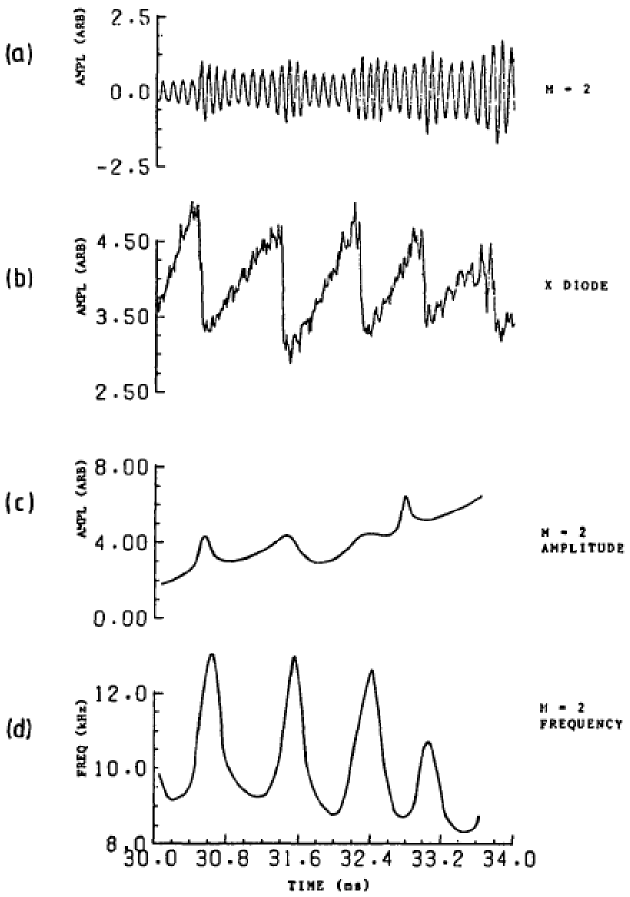


Figure 23: Response of the $m = 2$ poloidal magnetic signal amplitude and frequency to the sawtooth instability. Figure taken from [15].

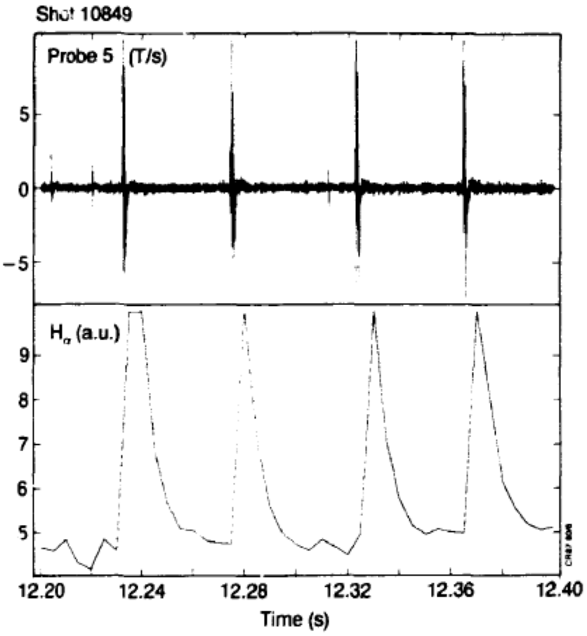


Figure 24: ELM-like activity observed from a magnetic probe and corresponding soft X-ray emission from the wall (near the X-point) following an H-mode. Figure taken from [18].

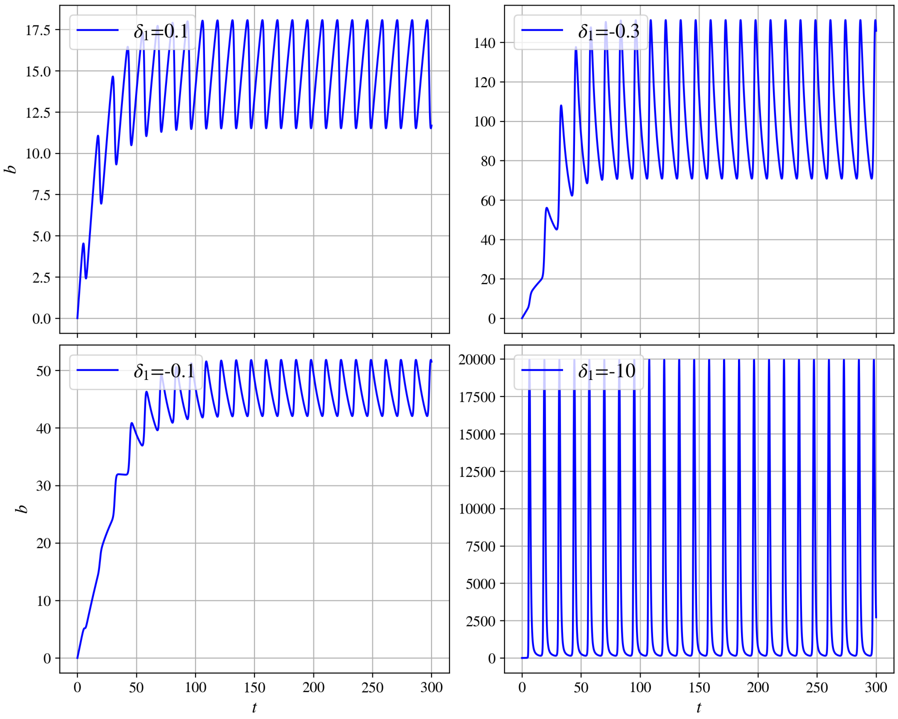


Figure 25: Behaviour of the ANAC model coupled to the equilibrium mode with varying δ_r. As δ_r is varied the oscillations become inverted and at large δ_r the behaviour resembles ELMs [71].

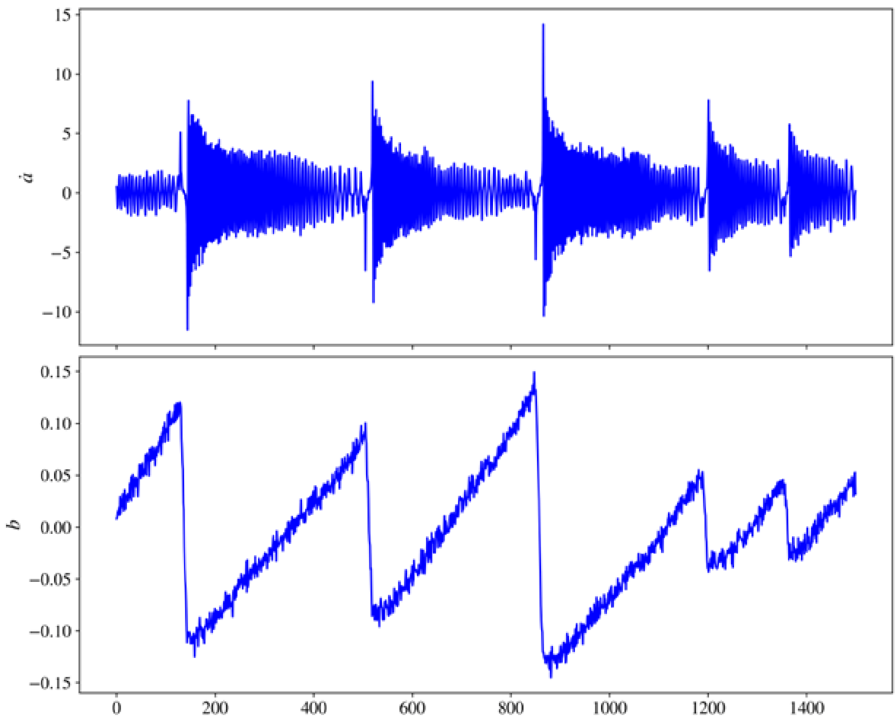


Figure 26: Simulation of the ANAET model with additive noise in the state variables and observational noise. Additive noise produces aperiodic oscillations in b and a noisy signal.

2.5 Conclusions

We have presented an overview of tokamak instabilities, the derivation of simplified bifurcation theory models and their qualitative similarity. We paid close attention to how the ANAET model could reproduce key aspects of the sawtoothing instability from relatively simple arguments. We then motivated that direct links can be made between the mode amplitudes of the ANAET model and experimental measurements. Explicitly, the mode amplitude $\dot{a}$ can be likened to Mirnov measurements and the mode amplitude b can be likened to soft X-ray measurements. In the following sections we shall discuss two data-driven approaches which can be used either to identify models of this form or attempt to match the ANAC or ANAET model directly to data.

3 SINDy as an approach for sparse equation recovery

In this section we review the sparse identification of nonlinear dynamics (SINDy) [97] as an approach for deriving ordinary differential equations from data. The application of SINDy is of interest as it aims to return sparse, interpretable dynamical systems which best describe the data. SINDy selects a small set of appropriate functions from a library of possible functions to generate models which fit the data. By enforcing sparsity during model selection, a combinatorically large selection of candidate models can be narrowed down to a handful of suitable models with fitted coefficients.

In this chapter, we will introduce the main optimisation method behind SINDy and discuss a broad selection of extensions relevant to this thesis. We will also discuss the general approach of model selection and validation with SINDy, and pay particular attention to extensions which help address weaknesses behind SINDy. Where relevant we also mention some extensions which could be explored in reference to possible future work.

3.1 SINDy

The field formally begins from the development of symbolic regression developed by refs. [57] and [74] which allows the physical form of governing equations to be distilled from data. Importantly, the number of terms appearing in the governing equations is balanced against predictive power allowing for comprehensible models. In contrast, methods such as deep neural networks provide future state predictions of dynamical systems, often following a black-box approach, typically focus on prediction and do not provide interpretable governing equations [123]. This symbolic regression method has more recently been developed into a sparse regression framework, SINDy [97], which is the main subject of this review.

The goal of SINDy is to learn the function $\boldsymbol{f}$ in the following type of dynamical systems

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{f}(\boldsymbol{x}(t)), \tag{3.1}$$

where $\boldsymbol{x}(t) \in \mathbb{R}^n$ is the state vector. In this form, $\boldsymbol{f}$ can represent ordinary differential equations such as the Lorenz equations. The key observation leveraged by this technique is that the function $\boldsymbol{f}$ is typically sparse in the space of possible functions which could appear on the right-hand side of equation (3.1). This observation allows for the solution of overdetermined systems, where the dimensions of the data typically far exceeds the number of unknowns to be determined. The sparsity constraint also allows for the discovery of interpretable models which avoid overfitting.

To find $\boldsymbol{f}$, the system given in equation (3.1) is framed as a linear system of form

$$\dot{\boldsymbol{x}} = \boldsymbol{\Theta}(\boldsymbol{x})\boldsymbol{\xi}, \tag{3.2}$$

with the goal to find the matrix $\boldsymbol{\Theta}$ containing possible features which are mapped to $\dot{\boldsymbol{x}}$ via a series of coefficients denoted $\boldsymbol{\xi}$. Time series collected at $t_1, t_2, \dots, t_m$ for states $x_1, x_2, \dots, x_n$ are arranged column-wise in a data matrix

$$\boldsymbol{X} = \text{time}\left\downarrow \overset{\xrightarrow{\qquad\qquad\text{state}\qquad\qquad}}{\begin{bmatrix} x_1(t_1) & x_2(t_1) & \dots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \dots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \dots & x_n(t_m) \end{bmatrix}}\right., \tag{3.3}$$

and similarly

$$\dot{\boldsymbol{X}} = \begin{bmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \dots & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \dots & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \dots & \dot{x}_n(t_m) \end{bmatrix}. \tag{3.4}$$

The matrix $\dot{\boldsymbol{X}}$ is usually not measured and instead is approximated using $\boldsymbol{X}$ with, for example, a finite difference method. The data matrix representing an approximation of $\boldsymbol{f}$ must then be formed. We construct the library $\boldsymbol{\Theta}(\boldsymbol{X}) \in \mathbb{R}^{m\times p}$ which contains the space of possible functions of size p. An example of this library is given for all possible polynomial terms up to quadratic nonlinearities

$$\boldsymbol{\Theta}(\boldsymbol{X}) = \begin{bmatrix} 1 & \boldsymbol{X} & \boldsymbol{X}^{P_2}, \end{bmatrix} \tag{3.5}$$

where polynomial libraries are denoted $\boldsymbol{X}^{P_2}$ for quadratic nonlinearities. Explicitly this is written as

$$\boldsymbol{X}^{P_2} = \begin{bmatrix} x_1^2(t_1) & x_1(t_1)x_2(t_1) & \dots & x_2^2(t_1) & \dots & x_n^2(t_1) \\ x_1^2(t_2) & x_1(t_2)x_2(t_2) & \dots & x_2^2(t_2) & \dots & x_n^2(t_2) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_1^2(t_m) & x_1(t_m)x_2(t_m) & \dots & x_2^2(t_m) & \dots & x_n^2(t_m) \end{bmatrix}. \tag{3.6}$$

The aim is then to find the sparse matrix of coefficients $\boldsymbol{\xi} = [\boldsymbol{\xi}_1 \ \boldsymbol{\xi}_2 \ \dots \ \boldsymbol{\xi}_n]$ which gives the active functions from the space of possible functions. Each column vector $\boldsymbol{\xi}$ corresponds to one state variable. The problem is then written as

$$\dot{\boldsymbol{X}} = \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}. \tag{3.7}$$

All that remains is to solve this overdetermined system of equations to determine Ξ. The simplest way to view this problem is to define the error between the prediction of the right-hand side, $\boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}$, depending on some choice of $\boldsymbol{\xi}$, and the observed data matrix $\dot{\boldsymbol{X}}$

$$E_2(\boldsymbol{\xi}) = ||\dot{\boldsymbol{X}} - \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}||_2^2. \tag{3.8}$$

Here the norm $||\cdot||_2^2$ denotes what is known as the least-squares solution, see for example ref [133][pg. 260]. Our error E then depends on the choice of $\boldsymbol{\xi}$. We are then interested in minimising the error, or the difference between the prediction and the observed data. Explicitly we can write this for each time observation in a column of $\dot{\boldsymbol{X}}$ as

$$\min_{\boldsymbol{\xi}} \frac{1}{m}\sum_{k=1}^{m}(\dot{\boldsymbol{X}}_{kl} - (\boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi})_{kl})^2 = \min_{\boldsymbol{\xi}} \frac{1}{m}||\dot{\boldsymbol{X}} - \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}||_2^2, \tag{3.9}$$

so for each column l (or state vector) we minimise the residual sum of squares between the prediction and the observed points in the time series. However, we will see that it is important to add what are called penalisation terms to this solution which may control factors such as how many elements appear in $\boldsymbol{\xi}$. If we include penalisation terms, the goal is to find $\boldsymbol{\xi}$ such that the following is minimised

$$\min_{\boldsymbol{\xi}} ||\dot{\boldsymbol{X}} - \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}||_2 + \lambda_1||\boldsymbol{\xi}||_1 + \alpha||\boldsymbol{\xi}||_2, \tag{3.10}$$

where $||\cdot||_1$ is the ℓ_1 norm, and the parameters λ_1 and α control the strength of the penalisation terms. The ℓ_1 norm is defined

$$\ell_1 = ||\boldsymbol{\xi}||_1 = \frac{1}{p}\sum_{k=1}^{p}|\xi_p|.$$

When $\alpha = 0$, Equation (3.10) is known as LASSO regression. Similarly when $\lambda_1 = 0$, it is called Tikhonov regularisation [119] [pg. 130]. An important aspect of regularisation is the ability to promote sparsity by minimising the number of elements which appears in $\boldsymbol{\xi}$. It can be seen that an increasing number of large non-zero terms causes a larger contribution from the terms in λ_1 and α. As we are attempting to minimises the stated objective, this contribution from the magnitude of the coefficients therefore aids to promote sparsity. It is worth remembering that SINDy in this form cannot handle implicit ODEs. If this is relevant the method must be changed as in [99], however, this adds significant complexity to the method. Extensions to PDEs can also be considered, described in [108, 109] but with modern applications it is almost always recommended to use the weak-from for PDE identification, which will be described later.

While the least squares approach can be used, it can be sensitive to noise added to observations. The addition of any form of noise and numerical differentiation error to the data $\boldsymbol{X}$ means that the least squares problem can be ill-conditioned [109]. SINDy is effectively a regularised least-squares regression problem, though there is one distinction. In a linear regression we usually assume a relationship

$$y = f(x) + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \delta_\sigma^2)$ is Gaussian measurement noise with variance δ_σ^2 and $f(x)$ is a function which maps variables x to the observation y [133]. This is subtly different to SINDy, where if noise is added to the state variables the problem becomes

$$y = f(x + \epsilon).$$

The difference is that in a typical linear regression we assume that the features of the regression have additive noise. Within SINDy, if the measurable states are noisy then when polynomial terms are formed this noise can be amplified [170]. If the assumption of additive Gaussian noise is violated (non-Gaussian noise is added), this method will likely have poor performance [132]. This is because a least-squares solution assumes Gaussian noise, and a missing feature can be viewed as non-Gaussian noise.

3.1.1 STLSQ Method

```
import numpy as np
#solve the initial least squares problem
Xi = np.linalg.lstsq(Theta, Xdot)

#for the number of threshold passes
for i in range(10):
    smallinds=np.where(Xi<Lambda) #find the small indexes
    Xi[smallinds]=0 #set coefficients below lambda to zero

    #solve the least squares solution on the remaining coefficients
    Xi[~smallinds,:]=np.linalg.lstsq(Theta[:,~smallinds],Xdot[~smallinds, :])
```

Listing 1: Implementation of the sequential thresholded least squares algorithm used by [97].

The constants λ_1, α are known as hyper-parameters or learning parameters in machine learning. An optimal value must chosen so that the resulting model predicts accurately, and this optimal value tends to be specific to the given data-set at hand. Further, the optimal value of these hyper-parameters is often not clear and instead we are forced to use rules of thumb or model selection techniques to select them. Model selection techniques have become a staple for validating machine learning methods, and will described in the following sections. The guiding philosophy for SINDy is Occam's razor: the simplest model successfully describing the data should be favoured over more complex ones.

In the original publication by [97], an optimisation method termed sequential thresholded least squares (STLSQ) is used to find sparse solutions. In many ways, STLSQ is almost synonymous with SINDy and performs feature selection of the candidate library. This process involves solving an initial least squares solution, followed by explicitly setting all coefficients below a magnitude λ to zero, and then regressing on to the remaining terms. This process is then repeating until no new terms are thresholded between iterations, or a maximum number of iterations is performed. An outline of the code written in Python is shown in Listing 1, and we emphasise that λ is distinct from λ_1. In this thesis, we will only consider the threshold parameter denoted λ and not the LASSO regularisation denoted λ_1. While the performance of STLSQ is generally quite good [163] compared to other optimisers, one issue is immediately apparent: there is no reason not to expect small coefficient terms in the resulting equations. Thresholding terms based on their coefficient magnitude assumes that coefficients of active terms are all of similar magnitude and larger than λ.

Ultimately, the efficiency of the SINDy framework is largely dictated by four aspects [113]:

i. the dimension n of the state vector $\boldsymbol{x}(t)$ of the system and the number of measurements m in time,

ii. the number of terms, p, in the candidate library $\boldsymbol{\Theta}$,

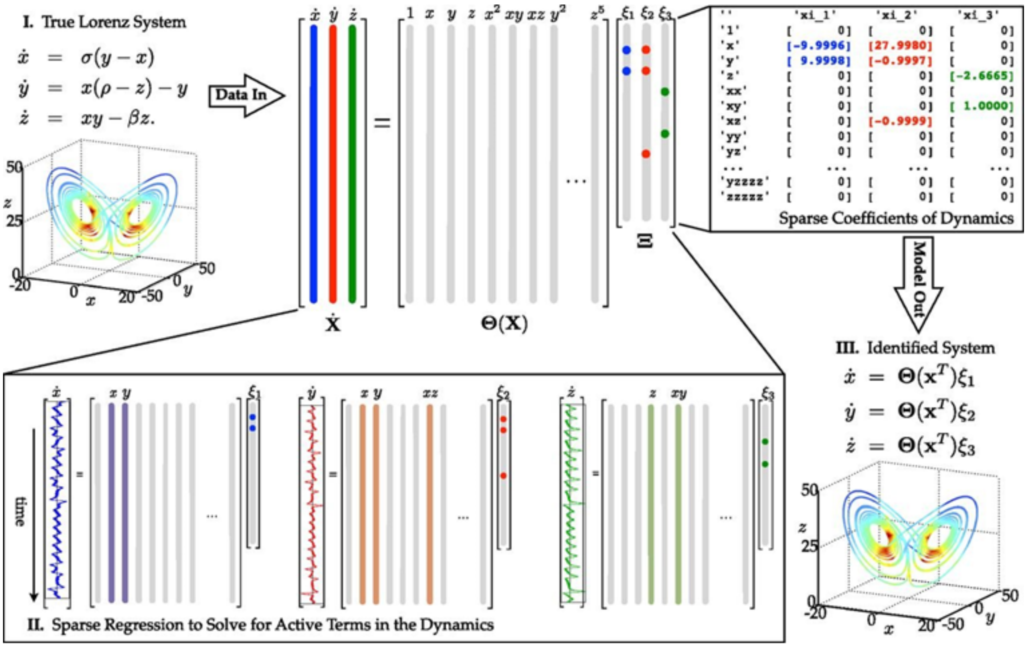
iii. the optimization method implemented,

iv. the quality of the data itself (for example sampling rates or noise).

The method outlined here has recently been created as a Python package called pysindy which can be easily imported through conda-forge [132]. This Python implementation allows for standardization of results and further several different optimizers are included to solve equation (3.7). While code was originally developed to perform the SINDy method, pysindy will be used throughout this thesis as it provides a reproducible framework with large documentation. One important distinction between the STLSQ algorithm included in pysindy and the original is that, following the final iteration of thresholding, an unregularised regression is performed with $\alpha = 0$ on the remaining active features. This is called an unbiasing step in the documentation, as non-zero regularisation tends to result in coefficient values which are biased towards the sparsity constraint.

The SINDy method represents a more recent advancement in sparse dynamical model discovery, with close relationships to nonlinear autoregressive moving average with exogenous inputs (NARMAX) methods (see ref [92]) and Bayesian regression formulations discussed by [49]. The main difference between SINDy and NARMAX methods lies in the choice of optimiser.

We finally present a classic schematic overview of the SINDy approach, shown in Figure 27. In this example we consider the reconstruction of a symbolic system of equations from purely time series data of the Lorenz system. We begin by generating input data from the Lorenz system for x, y and z which is then used to constructed the matrix $\dot{\boldsymbol{X}}$ using finite differencing. The regression problem is then performed by constructing a candidate feature library $\boldsymbol{\Theta}$ of terms that could appear in the resulting set of equations (such as polynomial terms of the input time series). The role of SINDy is to select the appropriate coefficients to map the library terms to the derivatives. For this step, any optimiser can be used but STLSQ is the most common. SINDy then identifies a small number of non-zero coefficients from the candidate feature library from which a symbolic set of equations can be formed. In the following section we will give a worked example.



I. True Lorenz System
$\dot{x} = \sigma(y - x)$
$\dot{y} = x(\rho - z) - y$
$\dot{z} = xy - \beta z.$
Data In
$\dot{x}$ $\dot{y}$ $\dot{z}$
1 x y z x^2 xy xz y^2 z^5
ξ_1 ξ_2 ξ_3
$\dot{\mathbf{X}}$
$\boldsymbol{\Theta}(\mathbf{X})$
Ξ
Sparse Coefficients of Dynamics
Model Out
III. Identified System
$\dot{x} = \boldsymbol{\Theta}(\mathbf{x}^T)\xi_1$
$\dot{y} = \boldsymbol{\Theta}(\mathbf{x}^T)\xi_2$
$\dot{z} = \boldsymbol{\Theta}(\mathbf{x}^T)\xi_3$
time
x y
x y xz
z xy
II. Sparse Regression to Solve for Active Terms in the Dynamics
50
25
0
-20
20
-50
x
y
z

```
''          'xi_1'      'xi_2'      'xi_3'
'1'      [      0]  [      0]  [      0]
'x'      [-9.9996]  [27.9980]  [      0]
'y'      [ 9.9998]  [-0.9997]  [      0]
'z'      [      0]  [      0]  [-2.6665]
'xx'     [      0]  [      0]  [      0]
'xy'     [      0]  [      0]  [ 1.0000]
'xz'     [      0]  [-0.9999]  [      0]
'yy'     [      0]  [      0]  [      0]
'yz'     [      0]  [      0]  [      0]
...          ...        ...        ...
'yzzzz'  [      0]  [      0]  [      0]
'zzzzz'  [      0]  [      0]  [      0]
```

Figure 27: Classic schematic of the SINDy algorithm. A set of time series data is formed into a regression problem by constructed a candidate feature library. SINDy then performs selection of the appropriate coefficients in the columns of ξ from which a set of symbolic equations is constructed. Figure taken from [97].

3.1.2 Example - Lorenz System

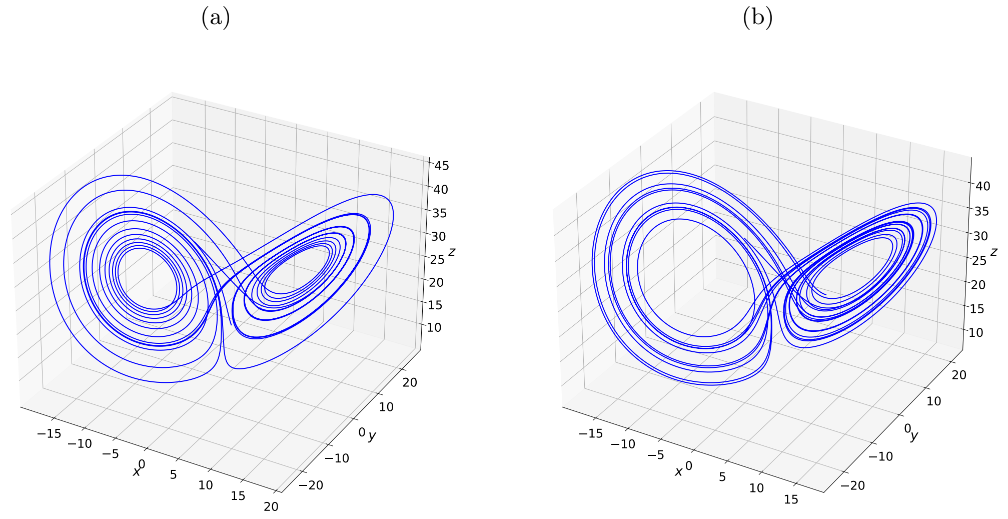


Figure 28: a) shows the Lorenz attractor of the original system integrated between $t = 0$ and $t = 20$. b) shows a comparison of the attractor dynamics of the identified system integrated over the same time range.

To illustrate SINDy and STLSQ, we use an example of the Lorenz equations as considered in the original SINDy paper [97]. The results here have been reproduced independently using a package known as pySINDy [132]. We explore the nonlinear coupled set of ODEs

$$\dot{x} = \sigma(y - x), \tag{3.11a}$$

$$\dot{y} = x(\rho - z) - y, \tag{3.11b}$$

$$\dot{z} = xy - \beta z, \tag{3.11c}$$

where $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$ with initial conditions $[x_0, y_0, z_0] = [-8, 7, 27]$. The Lorenz equations represent a good starting point, as they themselves are a reduced-order model deriving from Rayleigh-Bénard convection. To form our data matrices $\boldsymbol{X}$, we must gather a time-series measurement of the states x, y and z. The time-series are created by integrating the Lorenz system from $t = 0$ to $t = 100$ with time-step $\Delta t = 0.001$. The data matrices $\boldsymbol{X}$ and $\dot{\boldsymbol{X}}$ for the Lorenz system are filled with the time-series measurements

$$\boldsymbol{X} = \begin{bmatrix} x(t_1) & y(t_1) & z(t_1) \\ x(t_2) & y(t_2) & z(t_2) \\ \vdots & \vdots & \vdots \\ x(t_m) & y(t_m) & z(t_m) \end{bmatrix}, \qquad \text{and} \qquad \dot{\boldsymbol{X}} = \begin{bmatrix} \dot{x}(t_1) & \dot{y}(t_1) & \dot{z}(t_1) \\ \dot{x}(t_2) & \dot{y}(t_2) & \dot{z}(t_2) \\ \vdots & \vdots & \vdots \\ \dot{x}(t_m) & \dot{y}(t_m) & \dot{z}(t_m) \end{bmatrix},$$

where $t_1 = 0$, $t_2 = 0.001$ and so on. The data used to construct these matrices is referred to as the training data. The data matrix $\dot{\boldsymbol{X}}$ is found by using a second order finite difference scheme to approximate the derivatives. We can then consider construct a space of possible functions $\boldsymbol{\Theta}$ for this example. Here we will consider a library of polynomial terms up to and including third order, so explicitly takes the form

$$\boldsymbol{\Theta} = \begin{bmatrix} 1 & x(t_1) & y(t_1) & z(t_1) & x(t_1)^2 & \dots & z(t_1)^3 \\ 1 & x(t_2) & y(t_2) & z(t_2) & x(t_2)^2 & \dots & z(t_2)^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x(t_m) & y(t_m) & z(t_m) & x(t_m)^2 & \dots & z(t_m)^3 \end{bmatrix}.$$

We can now solve the regression problem

$$\dot{\boldsymbol{X}} = \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}$$

to determine the coefficients $\boldsymbol{\xi}$. The non-zero coefficients in the matrix $\boldsymbol{\xi}$ will then determine which terms appear in the governing equations. We implement STLSQ, used by [97], with a threshold parameter $\lambda = 0.1$. The identified model is

$$\begin{aligned} x &= -9.99978x + 9.99978y, \\ y &= 27.99802x - 0.9996y - 0.99994xz, \\ z &= -2.66659z + 0.99997xy. \end{aligned}$$

So we can see that from time-series generated from the ODE, we are able to reproduce the equations which generated them to a high degree of accuracy. Figure 28 shows the attractors for the original system and the identified set of equations. The identified equations are successfully able to capture the attractor dynamics and events such as lobe-switching of the original equations. Subsequent to this calculation, a full reproduction of the SINDy paper is now available in the pysindy documentation [132].

3.1.3 Choosing the threshold parameter

Choosing a model threshold λ must be done so with care. If the threshold is chosen too small, then many terms can appear in the resulting model. While it may initially seem beneficial to include as many terms as possible in the fitting process, this is actually dangerous and leads to a problem called over-fitting. This is especially true when there are noisy time-series. Additional terms allow the model too many degrees of freedom, and the resulting model will be a poor predictor. One way to avoid this is to include regularisation, for instance the parameter λ. For the following example, randomly distributed Gaussian noise is added to the data matrix $\boldsymbol{X}$ with zero mean and standard deviation 1×10^{-2}. The data matrix $\dot{\boldsymbol{X}}$ is then calculated from this noisy data using finite differences. No filtering method is applied to the noisy data.

Figure 29 shows the effect of varying the size of the threshold parameter over a range of values. The left of the figure shows the x equation integrated in time (red) compared to the noisy training data shown in black. The right column shows a bar chart of the number of non-zero coefficients appearing in the identified x equation (blue) and the true x equation outlined in red. We can see that the sparsity of the resulting model is controlled by the magnitude of λ. As λ is increased, the model becomes increasingly sparse and there are fewer non-zero coefficients. By $\lambda = 0.1$, we reach a solution which is close to the original set of equations. The time series shows that the best agreement is achieved by this model. This represents the idea of over-fitting, and the ideal of enforcing sparsity in the resulting models.

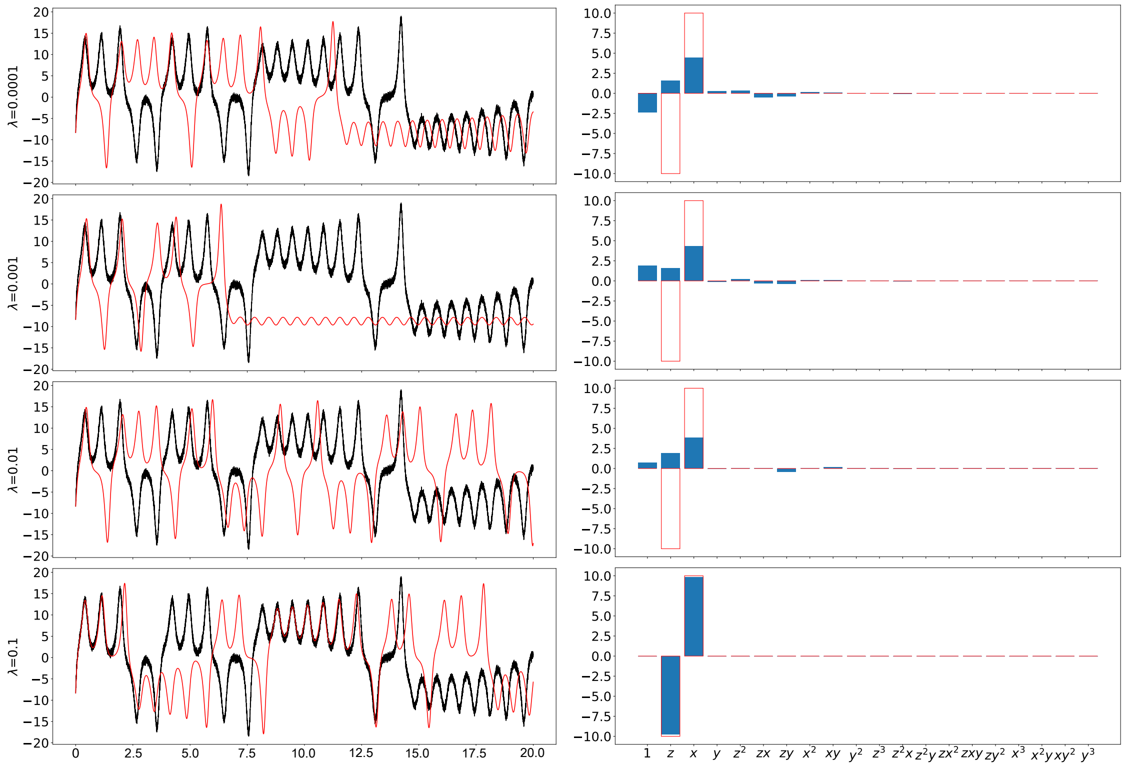


Figure 29: The effect of varying λ with the resulting model. The left column shows a time-series of the identified model (red) and the true model (black) integrated forward in time. The right column shows the terms appearing in the x equations. The blue bars are the identified model and the red outline are the actual parameters.

3.1.4 Choosing the library of possible functions

One of the assumptions in SINDy is that for the given input time series there should be a model in the library which allows for sparse representation of the data [122]. If a sparse model $\boldsymbol{f}$ is to be found, then the space of possible functions in equation (3.5) must contain a sparse basis for the dynamics. Ideally, many functions should be included in the feature library so that this is true. However, a large function space quickly leads to ill-conditioned problems as there is too much freedom to choose functions which fit the data.

Problems considered by [97] show that exclusion of the correct fitting terms returns ODEs with incorrect dynamics which are not sparse. In this way the method is advantageous as it provides a clear marker for when the sparse regression methodology has failed. Unfortunately, this is not the only instance whereby resulting models are not sparse and so only indicates that the approach has failed. The recommended approach to building the feature library is to start from a minimal low-order library and gradually increase the number of terms until a sparse model is obtained. Further, if knowledge of the system is known, then this can be used to guide the choice of terms.

However, it is not always required that the exact form of the governing equations is included in the feature library. In some cases, it is possible for SINDy to provide approximations to the dynamics it is attempting to identify, e.g., trigonometric functions by polynomials [97]. This has been used to identify Poincaré maps where analytic maps are not simple polynomials [129]. While an exact explicit recovery of the true map may not be possible, SINDy can still reproduce qualitative behaviour of the original maps, such as fixed points and stability. As such, there is not necessarily a general approach for choosing the library and this may become problem dependent.

Further, as SINDy rests on the assumption that the dynamical system governing the data is sparse, it relies on the choice of candidate functions allowing for a sparse representation. This also requires that measurements have been taken in such a way to allow for a sparse representation. If we instead measure combinations of the sparse basis features, the model recovered may no longer be sparse. Extensions to SINDy by [122] implemented a methodology which emphasised both the discovery of a coordinate system and a sparse dynamical model. While there are existing methods for low-dimensional representation of data, they either do not necessarily provide the correct basis for sparse representations of models, or cannot completely capture nonlinearities. An extra step can be included in SINDy where an autoencoder is used to find intrinsic coordinates which can support a sparse dynamical model.

3.1.5 Bifurcations

While the dynamical system given in equation (3.1) represents a large class of problems, often we are additionally interested in a bifurcation parameter μ of the system. The system is extended to include an equation representing the "dynamics" of the bifurcation parameter

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x};\mu), \tag{3.12a}$$

$$\dot{\mu} = 0 \tag{3.12b}$$

where μ is treated as a variable which is constant in time. μ can then be treated the same way as any other possible function in the library $\boldsymbol{\Theta}(\boldsymbol{X};\mu)$. Indeed, this was successfully applied in the original work of [97] to the logistic map and two-dimensional Hopf normal form. Further work has used birfucations to identify predator-prey type models of convection in magnetically confined plasmas [104].

3.2 Model Selection and Information Criterion

A key-stone of machine learning methods is model selection techniques. Model selection techniques aim to provide a robust method in which to choose data-sets and learning parameters appropriately. Choosing the data on which a machine learning model is trained on must be done with care. Many of the validation techniques outlined for NARMAX methods are also applicable to SINDy, see for example [92]. Within SINDy model evaluation metrics typically fall into one of two approaches:

1. We evaluate the predicted derivatives against the provided or estimated derivatives.
2. We integrate the resulting model and evaluate the trajectories against the provided trajectories.

The distinction between these approaches is important, and not always made explicit. The first approach effectively measures the quality of model fit in a graph fitting sense. We can use the training trajectories in our resulting model to get predictions of the derivatives which can be compared to the provided derivatives. This approach is the simplest, as it is fast to evaluate and does not require integration of the resulting model. However, there is no guarantee that a model with a good prediction of the derivatives does not become unstable in finite time [163]. The second approach of assessing performance requires integrating the resulting model itself. This has some advantages in that some assessment of the integrability of the model is performed, but is naturally more expensive to evaluate. Further, for chaotic models this approach can be harder to evaluate as trajectories will inevitably diverge. Even minor phase differences in oscillatory models can present issues in this approach and so it is rarely used to assess model performance. It is of course worth remembering that SINDy minimises the regularised least-squares error of the predicted derivatives to observed derivatives, and hence all models are generated based on this minimisation anyway.

Cross-validation

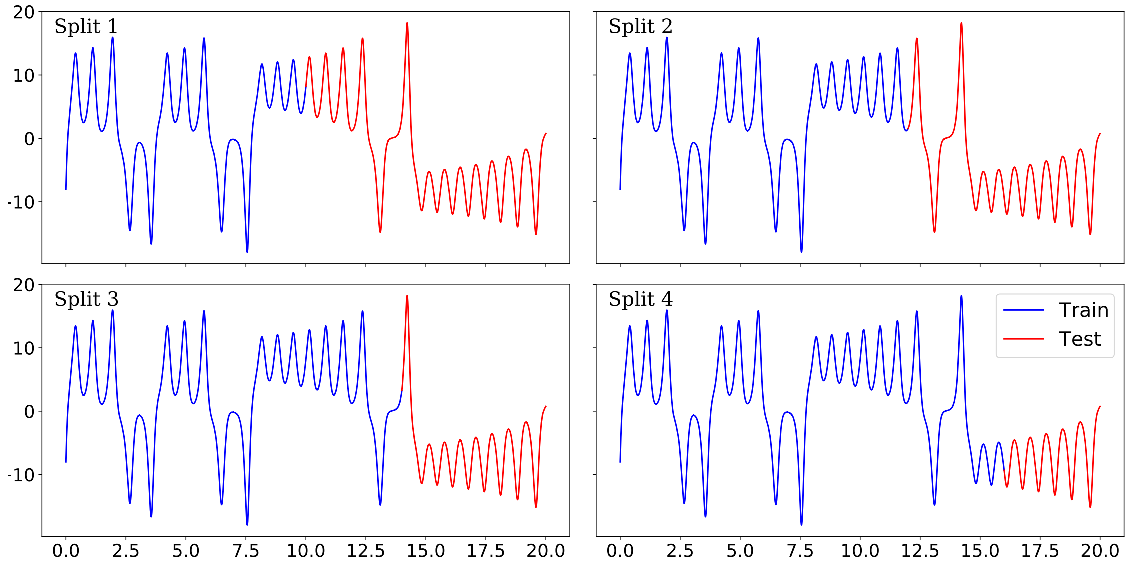


Figure 30: Illustration of time series split cross-validation for the x Lorenz equation. At each split, a larger training set (blue) is used and tested against the red trajectory.

One of the most commonly implemented model selection methods is known as cross-validation. This involves splitting data into a set which is used to train the model, and a set to test it. Several different kinds of cross-validation methods can be used, but some care must be taken when dealing with time-series. When finding models using SINDy, we are interested in the predictive capability of these models and how they forecast in the future. Time-series cross-validation can be performed by splitting the data at a number of times into a test and train set, where for each split, the train set gets progressively longer in time, as shown in Figure 30. For each split, a model is trained on the blue trajectory. The identified model can then be simulated forward in time and the mean-squared error calculated on the red test set. This process is then repeated for number of pre-decided splits. Cross-validating this way is useful to ensure that the length of the training data does not appreciably alter the result. It is also used when selecting parameters such as λ. For each value of λ, an average cross-validation residual sum of squares (RSS) can be calculated. The value of λ which produces the lowest average RSS on the test data can then be used. This approach is effectively part of the ensembling SINDy package, where average models are constructed from bagging the data [156].

AIC

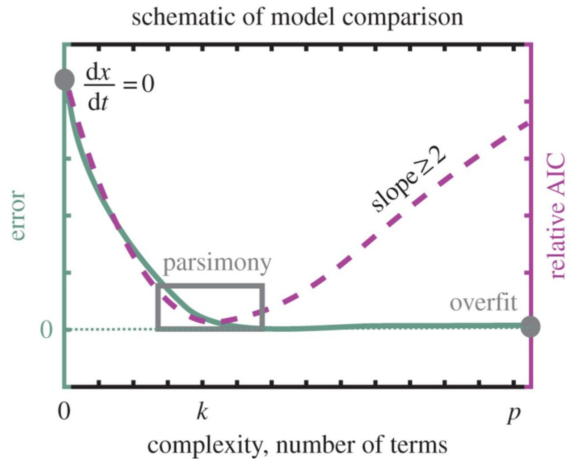


Figure 31: Schematic of the Pareto front showing the error on the left $y - axis$ in green and the relative AIC error on the right axis in purple. As the AIC score has a penalty of $2k$ the slope has gradient 2 past the elbow. Figure taken from [105].

When describing a set of data, we typically aim to describe the majority of the data with the simplest rule possible. This ethos describes the Pareto front analysis used when selecting the most suitable model from a number of candidates and is illustrated in Figure 31. The models with the lowest error for a set number of terms then define what is called the Pareto front [pg142][119]. As the number of terms increases, indicated by the green curve, we minimise the error. Beyond a certain sparsity, this reduction in error tends to plateau and indicates that the addition of further terms is fitting noise in the data. Beyond a given point, improving fit to the training is often at the expense of predictive capabilities of the model. The model will tend to fit noise in the training data which results in poorer predictions than a simpler model with fewer terms. With Pareto front analysis, the hope is that the error shows a sharp elbow with the inclusion of more terms in the model. This is the approach typically used with SINDy [97], but this elbow is not always clear [105].

One available method to automate model selection is the use of information criterion [119]. The Akaike information criterion (AIC) has already been used successfully in model selection with SINDy [105, 127]. Once a subset of suitable models has been chosen using a regularized least squares solution, AIC can be applied to find the optimal model. The AIC is defined for a given model i as

$$AIC_i = 2k - 2\ln(L(\boldsymbol{x}, \hat{\nu})), \tag{3.13}$$

where $L(\boldsymbol{x}, \nu) = P(\boldsymbol{x}|\nu)$ is the likelihood function of observation $\boldsymbol{x}$ given a set of model parameters ν, k is the number of parameters used by the model and $\hat{\nu}$ the best estimate of the parameter values [105]. This score then decreases with improving fit to the training data, but features a lower bound which increases with the number of nonzero terms. An important point to note is that this is a relative model score and as such always finds a best model. This is shown schematically in Figure 31 by the purple curve, where addition of more terms beyond the plateau in error causes an increase in the AIC score. This process does not in itself guarantee the model is suitable, purely that it is the best of some selection of candidate models and so is a relative scoring method. A similar model selection criterion is the Bayesian information criterion (BIC) defined as

$$BIC = \log(n)K - 2\log[L(\boldsymbol{x}, \hat{\boldsymbol{\nu}})], \tag{3.14}$$

where n is the sample size. The main difference between the two is that the BIC has been shown to theoretically select the correct model if it is included in the candidates, provided a large enough set of data $\boldsymbol{X}$ [pg. 152], [119]. The AIC requires a correction for the finite sample sizes, given by

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2(k+1)(k+2)}{(m-k-2)}, \tag{3.15}$$

where m is the number of observations. In practise, the log-likelihood is taken to be the residual sum of squares (RSS) with $\mathrm{RSS} = \sum_{i=1}^{\rho}(y_i - \hat{y}_i(x_i;\mu))^2$ with y_i the observed outcomes, x_i the independent variables and $\hat{y}$ the estimated model [127]. This scoring procedure pairs well with SINDy, allowing for the possibility of automated model selection. As the AIC score can be arbitrarily negative, often the relative AIC score is taken instead

$$AIC_{rel} = AIC - AIC_{min}, \tag{3.16}$$

where AIC_{min} is the AIC score of the minimum scoring model. As such, one model always has a value of zero in the relative scoring.

Some issues are evident with this model score. In particular, the addition of noise can change the minimum of the solution to equation (3.2) such that the correct model no longer has strongest support. Further, as the score is validated on test data using the RSS, chaotic models are still likely to give large RSS values. Even small errors in the estimation of the coefficients can cause significant differences in estimations after a characteristic time-scale [105]. The authors showed that the number of cross-validations could also influence the model score, with insufficient cross-validation giving strong support to the incorrect model. It is important to note that the authors calculate the AIC score over several integrated trajectories for deterministic models and different initial conditions, rather than evaluating the derivatives as done by ref [163].

KL Divergence

The KL divergence estimates the similarity between two distributions, and provides an approach for comparing the distributions of either the predicted derivatives or the integrated trajectories [138]. Given the training set data $\boldsymbol{X}$, we can calculate the variance-covariance matrix denoted $\boldsymbol{\Sigma}$, where the i^{th} row and j^{th} column entry is given by

$$\Sigma(\boldsymbol{X},\boldsymbol{X})_{ij} = \frac{\sum_i^{i=n}\sum_j^{j=n}(X_i - \bar{X_j})(X_i - \bar{X_j})}{(n-1)}, \tag{3.17}$$

where X_i denotes the columns of $\boldsymbol{X}$ and the barred variables denote the mean. Once a model has been identified we can then calculate estimates of the states denoted $\hat{\boldsymbol{X}}$ by integrating the resulting model and then find the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$. The KL divergence for multivariate distributions is given by

$$KL(\boldsymbol{\Sigma}||\hat{\boldsymbol{\Sigma}}) = \frac{1}{2}\left(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) - n + \ln\frac{|\hat{\boldsymbol{\Sigma}}|}{|\boldsymbol{\Sigma}|}\right). \tag{3.18}$$

where n is the number of equations. Given that we are comparing variance-covariance matrices, the underlying assumption is that the compared distributions are Gaussian in deriving this specific multivariate form. Despite this, it has been used successfully for non Gaussian distributions with SINDy [138]. We also note that predicted derivatives can equally be used, as opposed to integrated the resulting models.

3.3 Noisy data and weak SINDy

In most cases, it is likely that the collected data will be contaminated with some degree of noise. Often this noise is assumed to be Gaussian with zero mean so that $\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\eta}$ gives the noisy data with $\boldsymbol{\eta}$ the Gaussian noise. Calculating the derivative $\dot{\boldsymbol{X}}$ then introduces significant errors as a result. For the case of PDE identification, even extremely minor amounts of noise cause deterioration of the regression [114].

Several different approaches can be used to mitigate noise. In the original work by [97], a total-variation filtering method is used and demonstrated to perform well on given data-sets [82]. More general comparisons of taking numerical derivatives from noisy data are given in ref [145]. Without appealing to more advanced implementations of SINDy, several improvements can be made from data alone. This can be gained from simply providing more data, multiple trajectories with off attractor dynamics and transients or including symmetry reflections of the data.

In certain cases, it has been shown that the dynamics may still be recovered from highly corrupt measurements [102]. Constraints such as symmetry constraints and energy preserving nonlinearities discussed by [138, 148], globally stable models [149] and other general equality and inequality constraints [160] indirectly improve noise robustness by reducing the potential model space. Bayesian formulations [170] also provide improved robustness to noise, particularly in the low data limit. Model ensembling approaches also generate more reliable estimates of coefficient values by bagging the available data [156].

One of the most robust approaches recasts SINDy as an integral formulation, also known as the weak form [110, 142]. The weak form relies on recasting the SINDy problem in integral form either by using the fundamental theorem of calculus [110] or by transferring the derivatives to test functions [142]. Ref. [110] provides the first formulation of the weak problem, while [142] discusses the performance of this formulation in the context of PDEs which can be generalised to ODEs.

The weak form we describe here is introduced in the context of sparse PDE recovery from data. The notation used here is of course still applicable to ODEs and is the form used in pysindy. In this case, for some constant coefficients c_n where $n = 0, 1, ..., N$ we assume the PDE has the form

$$\partial_t \boldsymbol{u} = \sum_{n=0}^{N} c_n \boldsymbol{f}_n(\boldsymbol{u}, \partial_t \boldsymbol{u}, \partial_t^2 \boldsymbol{u}, \nabla \boldsymbol{u}, \nabla^2 \boldsymbol{u}, ...) = 0, \tag{3.19}$$

where $\boldsymbol{u}$ represents the state vector of the PDE. The problem is then cast into the weak formulation by multiplying by a differentiable test function $\boldsymbol{w}$ and integrating over some subset of the domain Ω_k defined by

$$\Omega_k = \{(x, y, z, t) : |x - x_k| \leq H_x, |y - y_k| \leq H_y, |z - z_k| \leq H_z, |t - t_k| \leq H_t\}, \tag{3.20}$$

such that the volumes are centred around randomly selected points (x_k, y_k, z_k, t_k) in the computational domain. H_x, H_y, H_z and H_t then represent the total size of the integration window in the x, y, z and t domains respectively. This process is then repeated a total of K different times for random choices of sub-domain locations so that the regression problem can be written as

$$\boldsymbol{q}_0 = \sum_{n=1}^{N} c_n \boldsymbol{q}_n = Q\boldsymbol{c} \tag{3.21}$$

where $Q = [\boldsymbol{q}_1, ..., \boldsymbol{q}_N]$ is the collection of column vectors $\boldsymbol{q}_n \in \mathbb{R}^K$ of the N different features of the PDE where each entry has the form

$$q_n^k = \int_{\Omega_k} \boldsymbol{w} \cdot \boldsymbol{f}_n \mathrm{d}\Omega_k, \tag{3.22}$$

and

$$q_0^k = \int_{\Omega_k} \boldsymbol{w} \cdot \partial_t \boldsymbol{u} \mathrm{d}\Omega_k. \tag{3.23}$$

We can then perform PDE identification by assuming the form of the PDE is not known, and substituting a selection of candidate functions for $\boldsymbol{f}_n$. This then casts Equation 3.21 as a linear regression problem, which minimises the residual sum of squared errors of the K different integrals on each sub-domain. Explicitly, the solution to Equation 3.21 is found by taking the pseudo-inverse

$$\tilde{\boldsymbol{c}} = Q^\dagger \boldsymbol{q}_0, \tag{3.24}$$

where $\tilde{\boldsymbol{c}}$ is the estimate of $\boldsymbol{c}$ given by the pseudo-inverse denoted by $Q^\dagger$. Then to implement STLSQ we can again iteratively threshold coefficients of $\tilde{\boldsymbol{c}}$ below a given magnitude and repeat the solution process. The power of the weak form lies in the ability to use integration by parts to transfer derivatives from the library functions to the test functions. For example we can write

$$q_0^k = \int_{\Omega_k} \boldsymbol{w} \cdot \partial_t \boldsymbol{u} \mathrm{d}\Omega_k = -\int_{\Omega_k} \partial_t \boldsymbol{\omega} \cdot \boldsymbol{u} \mathrm{d}\Omega_k, \tag{3.25}$$

provided the test function is zero on the boundary. This greatly improves the robustness to noise, as the test function is chosen to be smooth and continuously differentiable.

The weak form has been noted to provide substantial improvements with noise robustness, see for example refs [128, 171]. The primary limitation of this approach comes from the formation of integral windows over the data. In cases where we have periodic signals, if the integration domain spans multiple periods of that signal key features can be averaged over and in this sense can act like a low-pass filter [152]. Visualisation of this in the low data limit is given in ref [170]. An advancement of the weak form distinct from SINDy is called the sparse physics-informed discovery of empirical relations (SPIDER) [171] which augments the library of possible PDE features. By making use of rotational and translational invariance properties of vectors, substantial reductions in library terms can be made resulting in significant improvements.

3.3.1 Constraints

Included constraints in SINDy can be an important part of robust model identification. For example, the Lorenz equations exhibit the symmetry $(x, y, z) \to (-x, -y, z)$ and so any identified model must remain invariant to this transformation [138]. If we consider a general second order library, the regression problem can be expressed as finding the unknown coefficients of the following system of equations

$$\begin{aligned}
\dot{x} &= \gamma_0 + \gamma_1 x + \gamma_2 y + \gamma_3 z + \gamma_4 x^2 + \gamma_5 y^2 + \gamma_6 z^2 + \gamma_7 xy + \gamma_8 xz + \gamma_9 yz,\\
\dot{y} &= \beta_0 + \beta_1 x + \beta_2 y + \beta_3 z + \beta_4 x^2 + \beta_5 y^2 + \beta_6 z^2 + \beta_7 xy + \beta_8 xz + \beta_9 yz,\\
\dot{z} &= \rho_0 + \rho_1 x + \rho_2 y + \rho_3 z + \rho_4 x^2 + \rho_5 y^2 + \rho_6 z^2 + \rho_7 xy + \rho_8 xz + \rho_9 yz,
\end{aligned}$$

where the unknown coefficients are given by $\gamma_i, \beta_i, \rho_i$ for $i \in 0, 1, \ldots 9$. Enforcing the invariance in the resulting model reduces the number of unknown coefficients to

$$\begin{aligned}
\dot{x} &= \gamma_1 x + \gamma_2 y + \gamma_8 xz + \gamma_9 yz, \\
\dot{y} &= \beta_1 x + \beta_2 y + \beta_8 xz + \beta_9 yz, \\
\dot{z} &= \rho_0 + \rho_3 z + \rho_4 x^2 + \rho_5 y^2 + \rho_6 z^2 + \rho_7 xy,
\end{aligned}$$

which represents a substantial reduction in library size. Hence, different prior information can be used to inform library selection. This improves both the predictive performance of the models and their robustness to noise. As such there are at least two optimisers in pysindy that allow for constraints called: constrained SR3 [116] and mixed-integer optimisation for sparse regression (MIOSR) [160]. An implementation of constrained SR3 which produces models which are energy-preserving is given by trapping SINDy [149]. Constrained SR3 implements constraints as part of the objective which is being minimized, and as such constraints do not need to be obeyed. Conversely, MIOSR includes constraints as hard constraints which must be obeyed exactly.

3.3.2 Mixed Integer optimisation sparse regression

A comparatively recent addition to the optimisation in SINDy is the MIOSR optimiser [160]. Here we discuss the basic formulism as this optimiser is more involved than STLSQ but has exceptional performance compared to other optimisers available with SINDy [163]. MIOSR solves the optimisation problem

$$\min_{\boldsymbol{\xi}, \boldsymbol{z}} \quad ||\dot{\boldsymbol{X}}_j - \boldsymbol{\Theta}(\boldsymbol{X})\boldsymbol{\xi}||_2^2 + \alpha||\boldsymbol{\xi}||_2^2, \tag{3.26}$$

$$\text{s.t.} \quad M_i^l z_i \leq \boldsymbol{\xi}_i \leq M_i^U z_i \qquad i = 1, \ldots, D, \tag{3.27}$$

$$\sum_{i=1}^{D} z_i \leq k_j, \tag{3.28}$$

$$\xi_i \in \mathbb{R}, \quad z_i \in \{0, 1\}, \qquad i = 1, \ldots, D, \tag{3.29}$$

where M_i^l, M_i^U are lower and upper bounds on the coefficients and k_j is the total sparsity (maximum number of non-zero coefficients) for the j^{th} equation and z_i effectively labels which terms are active in the regression. This is termed a mixed-integer optimisation problem as both inequality constraints exist and z_i is restricted to integer values. In general this makes the problem non-convex and challenging to solve, so instead linear programming methods using brand-and-bound approaches are used (see for example ref [6][pg 272]). In the pysindy implementation, solutions are performed using the optimisation package Gurobi [172]. The MIOSR optimiser can also solve the entire system of d equations by forming the block diagonal system

$$\min_{\boldsymbol{\xi}} \left|\left| \begin{bmatrix} \dot{\boldsymbol{X}}_1 \\ \dot{\boldsymbol{X}}_2 \\ \vdots \\ \dot{\boldsymbol{X}}_d \end{bmatrix} - \begin{bmatrix} \boldsymbol{\Theta} & 0 & \ldots & 0 \\ 0 & \boldsymbol{\Theta} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{\Theta} \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi}_1 \\ \boldsymbol{\xi}_2 \\ \vdots \\ \boldsymbol{\xi}_d \end{bmatrix} \right|\right|_2^2 . \tag{3.30}$$

This means that constraints between coefficients of the form $\boldsymbol{A}\bar{\boldsymbol{\xi}} \leq \boldsymbol{b}$ where $\bar{\boldsymbol{\xi}}$ is a Dd length vector of all coefficients can now be used. Here $\boldsymbol{A} \in \mathbb{R}^{l \times Dd}$ represents a matrix of l constraints between the D different features with d dimensions. Rows of this describe linear systems of equations between the different variables. $\boldsymbol{b} \in \mathbb{R}^l$ is then a vector of real-numbers representing, for example, upper bounds that the constrained terms must not exceed. When sparsity is enforced on each equation separately, it is called group sparsity. When sparsity is enforced on all equations at once (the total allowable number of non-zero coefficients across all equations) it is termed target sparsity.

For the end-user, MIOSR has several favourable qualities:

1. Sparsity is implemented as a hard constraint, rather than a penalty term in the objective. This is favourable when correlated features are an issue as is often the case with say polynomial features in the regression library [160],
2. Models are provably optimal, as upper and lower coefficient bounds converge as the approach effectively performs a combinatorically hard search over the feasible solution space,
3. Additional constraints are also hard, and so must be exactly obeyed.
4. The addition of constraints tends to improve computational performance, as the search space of viable models get smaller.

Compared to STLSQ which requires optimisation of α and λ, MIOSR specifies the total sparsity k and α. Specifying the total sparsity instead of a threshold now means that small coefficient terms which could be relevant are not lost during optimisation. In ref [160], stark improvements are found using MIOSR compared to other available optimisers. Comparisons of $6D$ systems fit using STLSQ and MIOSR show that MIOSR greatly outperforms STLSQ where there are correlated library features and STLSQ can take incorrect intermediate steps (once a library term has been thresholded, it can never be recovered). In all cases discussed in this thesis we limit attention to STLSQ or MIOSR. While other optimisers are available, the performance is typically weaker than these two [163].

3.3.3 Bayesian SINDy/ Bayesian regression

The presence of noise in our measurements will undoubtedly introduce some errors in the estimation of our models. Several separate sets of measurements for the same dynamics can produce different resulting models and this would lead us to find a way to quantify uncertainty in model coefficients. Further, when our measurements are scarce we may also wish our uncertainty to reflect this. In fact, Bayesian regression pre-dates SINDy, most notably as a MATLAB script called `SparseBayes` which is an implementation by [44]. More recently two prominent implementations of Bayesian-SINDy have appeared, namely [157] and [170], though the former was only recently added to pysindy despite its relatively earlier publication. The distinct advantage of a Bayesian formulation comes from a natural quantification of the uncertainty in model coefficients due to observational errors, limited data and also lists the probability of inclusion of each candidate function. Ensemble SINDy can also provide inclusion probabilities for candidate functions [156].

3.4 Applications of SINDy in plasma physics

SINDy and its variants have seen some application to MHD and related fields, with a general recent review of sparse regression in plasma physics given by [164]. One of the earliest applications of SINDy to a plasma physics problem was by ref [104] where predator-prey type models are constructed for L-H transition. Simulations are performed of a magnetically conducting fluid in which there is a pressure source, and SINDy models are fit to 3 computed flow variables. By fitting dynamical systems sequentially in different solution regimes, the simulations are parameterised by a predator-prey model through different bifurcating solutions (ranging from different steady solutions to oscillatory convection). More recently SINDy has been used to fit reduced models to scrape-off layer simulations of divertor flux by ref [165]. Reduced models are derived by considering the electron density at the outer midplane of the tokamak and the electron temperature at the outer divertor. Both linear and nonlinear models are constructed that are capable of prediction of the electron density and electron temperature at the selected locations, and a discussion is given on control applications for future application to experiment.

Integral formulations have seen applications to the kinetic Vlasov equations for recovery of the integral form of PDEs from noisy particle in cell simulations [128]. Application of numerical differentiation techniques with noise often require the selection of some parameter, such a polynomial degree, which can be difficult to choose without prior knowledge of the clean data. As such, incorrect selection results in poor coefficient estimation of the model parameters. However, the PDE integral formulation requires some selection of the volume over which terms are integrated which is likely to require some prior knowledge of the system. This would have to be chosen so that it smaller than characteristic variations within the system. The use of integral terms with the Vlasov equations is shown to reduce inferred coefficient error from $20 - 30\%$ to approximately 2%. The same approach was also used to extract the Hasegawa-Wakatani model from synthetic data generated from the equations themselves [153]. Important discussions are given on domain selection and total length of data with the integral form. Some care should be taken, however, as the work in refs [128, 153] compute the derivatives first before integrating.

3.5 SINDy with higher dimensional data

The size of the SINDy feature library grows rapidly with an increase in data dimensionality. As such SINDy is typically limited to analysis of a handful of time-series only. Many realistic applications might apply to higher dimensional data such as PDE simulations which can have many more degrees of freedom. If we hope to identify some system of ODEs from the PDE data, we must first appeal to a dimensionality reduction technique. A corner-stone of data reduction techniques is the singular-value decomposition (SVD). This method is key in being able to sparsely represent high-dimensional data-sets with low rank approximations. SVD also forms the backbone of many other model reduction techniques including dynamic mode decomposition (DMD) [101] and the Hankel Alternative View of Koopman theory (HAVOK) method [103].

3.5.1 Singular Value Decomposition

Suppose we have a dataset $\boldsymbol{X} \in \mathbb{C}^{n\times m}$ where again the rows of $\boldsymbol{X}$ could represent temporal measurements. As a result we typically have $n \gg m$. The singular value decomposition, SVD, of $\boldsymbol{X}$ can be written

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \tag{3.31}$$

where $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{C}^{m\times m}$ are unitary matrices with orthonormal columns and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times m}$ is a diagonal matrix [pg5][98]. The diagonal entries $\sigma_i = \Sigma_{ii}$ are called the singular values and are arranged in descending order of magnitude. The columns of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are called the left and right singular vectors respectively. SVD has two important applications: the calculation of the pseudo-inverse which can be used to find a least squares solution, and a rank r approximation to the matrix $\boldsymbol{X}$.

One of the most important applications of SVD is reducing the size of a matrix. SVD is a powerful tool which allows many of the following methods to work, and allows large data-sets to be represented by the product of 3 much smaller matrices. The rank r approximation involves truncating equation 3.31 at the leading r singular values to create an approximation of the matrix $\boldsymbol{X}$. As these singular values are organised in descending order, we can think of them as labelling the importance of each of the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. The truncated matrix $\tilde{\boldsymbol{X}}$ is given by sum of r rank-1 matrices

$$\tilde{\boldsymbol{X}} = \sum_{k=1}^{r} \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^* = \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^* + \cdots + \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^* \tag{3.32}$$

where $\boldsymbol{u}_k$ and $\boldsymbol{v}_k^*$ are the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ respectively. Thinking of a truncated SVD as the sum of columns of the two matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ is key to the intuition behind some the methods discussed in Koopman theory. The decomposition in equation (3.31) is that the columns of $\boldsymbol{U}$ form an orthonormal basis for the column space of $\boldsymbol{X}$ and the columns of $\boldsymbol{V}$ form an orthonormal basis for the row space of $\boldsymbol{X}$ [pg13][98]). If the matrix $\boldsymbol{X}$ has columns of spatial measurements at fixed times, then the columns of $\boldsymbol{U}$ represent the spatial patterns and the columns of $\boldsymbol{V}$ the temporal evolution. This idea can be used in proper orthogonal decomposition (POD) to find an orthonormal basis which reconstructs the time-dynamics of a system [35].

In fluids applications, the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$ (which turn out to be the square of the singular values) are related the to the kinetic energy of the fluid [35]. We can therefore think of the SVD as providing a decomposition which maximizes the energy reconstruction. More generally, an inner product is defined for MHD problems by [150] which again allows interpretation of the singular values as representing the total energy when there is also the influence of a magnetic field. Other relevant modal decompositions are, for example, DMD where the identified spatial modes evolve like $\exp(i\omega t)$ where $\omega = \omega_r + i\omega_i$ is a complex growth-rate. For applications with SINDy, this comes with the additional challenge of correlated modes [154].

3.5.2 Proper Orthogonal Decomposition

We now discuss proper orthogonal decomposition which aims to decompose a field by projecting onto a set of orthogonal basis modes. Consider a system of nonlinear PDEs written

$$\boldsymbol{u}_t = \boldsymbol{f}(\boldsymbol{u}, \boldsymbol{u}_x, \boldsymbol{u}_{xx}, \ldots, \boldsymbol{x}, t; \boldsymbol{\beta}), \tag{3.33}$$

where $\boldsymbol{\beta}$ are a set of parameters. If we assume an expansion of the solution $\boldsymbol{u}$ in terms of a set of optimal spatial modes ψ_k and corresponding amplitudes $\boldsymbol{a}_k$ we can write the separable solution of the form

$$\boldsymbol{u} = \sum_{k=1}^{n} \boldsymbol{a}_k(t)\psi_k(\boldsymbol{x}). \tag{3.34}$$

In index notation, this can be written for the state vector $\boldsymbol{u}$

$$u_j = \sum_{k=1}^{n} a_{jk}\psi_k(\boldsymbol{x}). \tag{3.35}$$

The POD expansion provides a data-driven approach to find an optimal set of basis modes for a flow (as opposed to a Fourier mode expansion in spectral methods for instance). To apply this method, we take snapshots of the flow at time t_k as $\boldsymbol{u}_k = [u(x_1, t_k) \quad u(x_2, t_k) \ \ldots \ u(x_n, t_k)]^T$ and arrange them in a large data matrix $\boldsymbol{X}$ as follows

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_m \end{bmatrix}, \tag{3.36}$$

for m different time measurements. We then perform the SVD on the matrix $\boldsymbol{X}$ as described above to find a set of orthonormal basis modes for the flow.

3.5.3 Dimensionality reduction with SINDy

One successful example of combining dimensionality reduction with SINDy is given by ref. [138]. The authors consider a modification of Rayleigh-Bénard convection in an annular domain. When heated from below, the fluid rotates in a clockwise and anti-clockwise directions. By applying DMD to the temperature and velocity fields separately, the authors are able to create a reduced order resembling the Lorenz equations which captures statistical properties of the flow such as rotation of the flow and dominant frequencies of rotation.

However, limitations of using DMD with SINDy are apparent as DMD identifies spatial modes which evolve as $\exp(\omega t)$ in time. As such, the only difference in the time series is the amplitude and frequency of oscillation and while they may be linearly uncorrelated with different frequencies, nonlinear prodcucts of different modes can be correlated. For model recovery, this creates a highly correlated library and a challenging identification problem. Work by [154] addresses this issue by considering two points: there are only a handful of driving DMD modes and the rest are harmonic combinations of the others and by using the statistical randomized dependence coefficient to measure nonlinear correlations. A reduced-model is then constructed from only 4 DMD modes of the original 16 which successfully reproduces the cavity flow.

SINDy has also been used with POD decomposition techniques, such as by ref [148]. In this application the authors perform POD of a chaotic electroconvection problem. Interestingly here POD of only one field of the full problem is performed, from which a simplified ODE system is constructed. Further, in the original SINDy publication the evolution of 3 POD modes are used to capture flow past a cylinder [97]. Other more advanced applications of SINDy have come about which use autoencoders to find reduced representations of equations [122]. This is achieved by introducing a complicated loss function which demands that the dimensionality reduction of the autoencoder provides a good reconstruction of the data, but also produces a sparse model in the SINDy regression. This work has been extended in a number of interesting directions, where Lorenz-like models are discovered from videos of chaotic waterwheels and partial measurements [159].

3.5.4 SINDy with partial measurements

A particular weakness of the SINDy method is the inability to predict models when measurements of a given variable have not been made. While autoencoders may be used to find the appropriate coordinate systems for mixed observations [122], other methods must be used when measurements are completely absent. One such example follows from DMD, called the Hankel alternative view of Koopman (HAVOK) method [103]. HAVOK relies on Koopman theory (see [Pg. 7][98]), where the goal is find a coordinate system which transform nonlinear dynamics into linear dynamics. Effectively this means using time-delay embeddings to attempt to reconstruct the attractor from partial measurements of the system, see for example [167]. Other approaches combine a delay-embedding with an autoencoder, as given in ref [159]. Other approaches simply use POD of a single PDE variable from which to construct models, as already discussed in ref [148]. In general though, these problems are challenging as often there is no guarantee that a chosen embedding admits a sparse solution. Several issues when using partial measurements are outlined by [175].

3.6 Conclusions

We have presented an introduction to SINDy as it was originally published in ref. [97] and discussed the limitations of this approach. Several extensions for SINDy have been discussed such as the weak form as an approach for handling noisy signals. We have also given an overview of SINDy applied with decomposition methods such as POD, with some relevant applications in plasma physics. In the following section we will apply SINDy to a system of ODEs which follow a close derivation to the Lorenz model, but also includes the influence of magnetic field. We will validate some test cases for SINDy using model selection techniques, and discuss some apparent challenges. We will primarily explore the weak form extension with these ODEs and assess the noise robustness and sensitivity to sampling rates.

4 SINDy with Magnetoconvection

In this section we introduce magnetoconvection theory and the derivation of a low-order model which qualitatively describes the PDE system [7]. The low-order model facilitates simpler exploration of the full phase-space of solutions and the applicability of the magnetoconvection problem studied by Chandrasekhar [2] in tokamaks is discussed by ref [80]. Parallels between the magnetoconvection problem with a horizontally imposed magnetic field and reduced MHD (RMHD) problems for tokamaks can be drawn where the buoyancy term driving convection is identified with a term related to curvature of the magnetic field in tokamaks. These equations are used in the numerical simulation of ELMs by [83], for example. Here we study the case of a vertically imposed magnetic field, as we are primarily interested in assessing the performance of SINDy, though similar reduced models can be found in the case of a horizontal field [11].

The construction of a 5^{th} order system [7] will be discussed below, which importantly allows a rich description of the solution behaviour. This system provides a suitable test-bed for SINDy as it displays a wide range of behaviours such as periodic, semi-periodic and chaotic solutions, the behaviour of which is outlined in ref [13]. Studying this simplified model provides insight into the bifurcation structure of the PDE system and illuminates cases where we expect nonlinear contributions to modify conclusions from linear theory.

4.1 Magnetoconvection in 2D

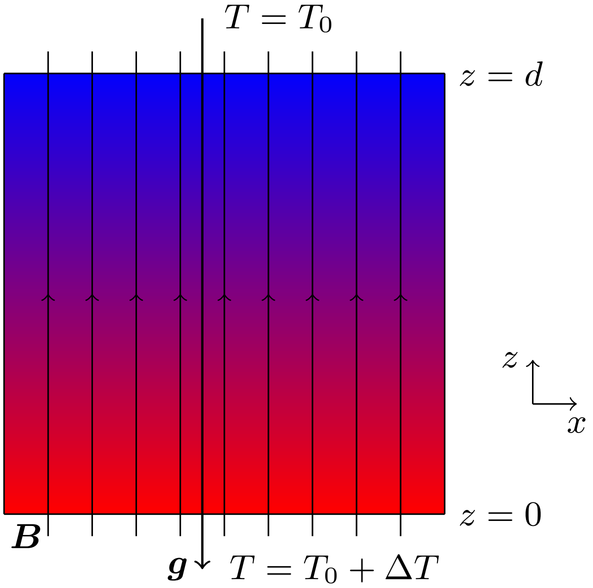


Figure 32: Configuration for the magnetoconvection problem of an electrically conducting fluid with an imposed vertical magnetic field.

We now discuss the derivation of linear and weakly nonlinear theory models for magnetoconvection, the details of which are adapted from ref. [2][pg. 146] and ref. [9]. The presence of a magnetic field in an electrically conducting fluid has important effects on the fluid stability. For one, Alfvén's frozen flux theorem states that the fluid will be constrained with the motion of the field lines in the ideal case. This implies that in general the effect of the Lorentz force will be to inhibit convection. Further, the presence of Alfvén waves suggests the possibility that the onset of convection can begin through overstable oscillations in the magnetic field. By using linear theory, we can confirm that these statements are possible.

We consider an electrically conducting fluid between two boundaries with normal in the $\hat{\boldsymbol{z}}$ direction with an applied uniform heating from below and magnetic field imposed in the vertical direction with value $\boldsymbol{B}_0 = B_0\hat{\boldsymbol{z}}$. The upper and lower boundaries are situated at $z = 0$, and $z = d$ respectively with the bottom boundary being held at temperature $T = T_0 + \Delta T$ and the top at $T = T_0$, shown in Figure 32. The equations describing thermal convection in a conducting fluid in the presence of a magnetic field are as follows

$$\begin{aligned}\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} &= -\frac{1}{\rho_0}\nabla p - \frac{\rho}{\rho_0}g\hat{\boldsymbol{z}} + \frac{1}{\mu_0\rho_0}(\nabla\times\boldsymbol{B})\times\boldsymbol{B} + \nu_f\nabla^2\boldsymbol{u}, &(4.1)\\ \frac{\partial T}{\partial t} &= -\nabla\cdot(T\boldsymbol{u}) + \kappa\nabla^2 T, &(4.2)\\ \frac{\partial \boldsymbol{B}}{\partial t} &= \nabla\times(\boldsymbol{u}\times\boldsymbol{B}) + \eta\nabla^2\boldsymbol{B}, &(4.3)\\ \nabla\cdot\boldsymbol{B} &= 0, &(4.4)\end{aligned}$$

where the additional magnetic force arises from the Lorentz force with a small effect from electric fields

$$\boldsymbol{F}_B = \frac{1}{\mu_0}\boldsymbol{j}\times\boldsymbol{B}, \tag{4.5}$$

and equation (4.3) is the induction-diffusion equation. The constants are defined as ν_f the viscous diffusivity, κ the thermal diffusivity, η the magnetic diffusivity and μ_0 the magnetic permeability. We assume the Boussinesq approximation holds valid so that $\nabla\cdot\boldsymbol{u} = 0$ and $\rho/\rho_0 = 1 - \alpha(T - T_0)$, where α is the coefficient of volume expansion. The basic state is given when $\boldsymbol{u} = 0$ and the temperature gradient is purely conductive. Equation 4.1 then becomes

$$0 = -\frac{\nabla p}{\rho_0} - g(1 - \alpha(T - T_0))\hat{\boldsymbol{z}}. \tag{4.6}$$

The pressure gradient is then at most only a function of z, e.g., $p(z)$ and acts to balance the temperature in the z-direction. As the situation is steady, we have from equation (4.2)

$$\nabla^2 T = 0. \tag{4.7}$$

In equilibrium, $T = T(z)$ only and hence we have that

$$\bar{T}(z) = T_0 - \frac{\Delta T z}{d}, \tag{4.8}$$

is the vertically averaged temperature profile after applying the temperature boundary conditions at $z = 0, d$. As the temperature gradient varies linearly in z in the base state, define $\beta = -\frac{\mathrm{d}\bar{T}}{\mathrm{d}z} = \frac{\Delta T}{d}$ so that

$$\bar{T}(z) = T_0 - \beta z. \tag{4.9}$$

To find the pressure use

$$\begin{aligned}\frac{1}{\rho_0}\frac{\partial p}{\partial z} &= -g(1 - \alpha(\bar{T} - T_0)) = -g(1 - \alpha - \beta z), &(4.10)\\ \implies \bar{p}(z) &= p_0 - \rho_0 g z(1 + \tfrac{1}{2}\alpha\beta z), &(4.11)\end{aligned}$$

where p_0 is the pressure at the centre. This is the full description of the basic state. We now expand equation (4.1) and equation (4.2) about the basic states

$$p = \bar{p} + \delta p, \qquad T = \bar{T} + \delta T, \tag{4.12}$$

to get

$$\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u} = -\frac{1}{\rho_0}\nabla \delta p + \alpha g \delta T \hat{\boldsymbol{z}} + \frac{1}{\mu_0}\rho_0(\nabla \times \boldsymbol{B}) \times \boldsymbol{B} + \nu_f \nabla^2 \boldsymbol{u}, \tag{4.13}$$

$$\tag{4.14}$$

for the momentum equation and

$$\frac{\partial \delta T}{\partial t} = -\nabla \cdot (\delta T \boldsymbol{u}) + \beta w + \kappa \nabla^2 \delta T, \tag{4.15}$$

for the energy equation. We next non-dimensionalise the problem by the thermal conduction time d^2/κ and layer depth d

$$\boldsymbol{x} = \tilde{\boldsymbol{x}} d, \qquad t = \tilde{t} d^2/\kappa, \qquad \boldsymbol{u} = \tilde{\boldsymbol{u}}\kappa/d, \qquad \boldsymbol{B} = \tilde{\boldsymbol{B}} B_0, \qquad \delta T = \Delta T \tilde{T} \tag{4.16}$$

and introduce the dimensionless numbers

$$Pr = \frac{\nu_f}{\kappa}, \qquad \zeta = \frac{\eta}{\kappa}, \qquad R = \frac{g\alpha \Delta T d^3}{\kappa \nu_f}, \qquad \mathcal{Q} = \frac{B_0^2 d^2}{\mu_0 \rho_0 \eta \nu_f} \tag{4.17}$$

where $\mathcal{Q}$ is the Chandrasekhar number, Pr and ζ are the two Prandtl numbers and R is the Rayleigh number. We now list the dimensionless equations, understanding that the non-tilde variables are now dimensionless

$$\frac{1}{Pr}\left(\frac{\partial \boldsymbol{u}}{\partial t} + (\boldsymbol{u} \cdot \nabla)\boldsymbol{u}\right) - \zeta Q(\boldsymbol{B} \cdot \nabla)\boldsymbol{B} = -\nabla\left(P_{term} + \frac{\zeta Q}{2}|\boldsymbol{B} \cdot \boldsymbol{B}|\right) + RT\hat{\boldsymbol{z}} + \nabla^2 \boldsymbol{u}, \tag{4.18}$$

$$\frac{\partial T}{\partial t} = -\nabla \cdot (T\boldsymbol{u}) + w + \nabla^2 T, \tag{4.19}$$

$$\frac{\partial \boldsymbol{B}}{\partial t} = \nabla \times (\boldsymbol{u} \times \boldsymbol{B}) + \zeta \nabla^2 \boldsymbol{B}, \tag{4.20}$$

$$\nabla \cdot \boldsymbol{u} = 0, \qquad \nabla \cdot \boldsymbol{B} = 0. \tag{4.21}$$

together with the solenoidal constraints for $\boldsymbol{u}$ and $\boldsymbol{B}$, and the Lorentz force has been expanded using a vector identity. The pressure scaling has not been chosen as we intend to remove this term by constructing the vorticity equation. We now consider the case of motion confined in $x-z$ plane where both the vertical and horizontal are free boundaries. In dimensionless terms $x \in [0, L_x]$ and $z \in [0, 1]$ is the extent of the domain. We eliminate the pressure from equation (4.18) by taking the curl once and introducing the stream and flux functions defined by

$$\boldsymbol{u} \equiv (u, 0, w) = \nabla \times \psi \hat{\boldsymbol{y}} = \left(-\frac{\partial \psi}{\partial z}, 0, \frac{\partial \psi}{\partial x}\right), \qquad \boldsymbol{B} \equiv (B_x, 0, B_z) = \nabla \times A\hat{\boldsymbol{y}} = \left(-\frac{\partial A}{\partial z}, 0, \frac{\partial A}{\partial x}\right), \tag{4.22}$$

so that the current density and vorticity can now be defined

$$\boldsymbol{j} = \nabla \times \boldsymbol{B} = (\partial_{xy}A, -\partial_{xx}A - \partial_{zz}A, \partial_{zy}A), \qquad \boldsymbol{\omega} = \nabla \times \boldsymbol{u} = (\partial_{xy}\psi, -\partial_{xx}\psi - \partial_{zz}\psi, \partial_{zy}\psi), \tag{4.23}$$

respectively. The full 2D problem can then be written

$$\frac{1}{Pr}\left(-\frac{\partial\nabla^2\psi}{\partial t} - J(\psi,\nabla^2\psi)\right) = -\nabla^4\psi - R\frac{\partial T}{\partial x} - \zeta Q J(A,\nabla^2 A), \tag{4.24}$$

$$\frac{\partial T}{\partial t} + J(\psi,T) = \nabla^2 T, \tag{4.25}$$

$$\frac{\partial A}{\partial t} + J(\psi,A) = \zeta\nabla^2 A, \tag{4.26}$$

where

$$J(\psi,\nabla^2\psi) = \frac{\partial\psi}{\partial x}\frac{\partial\nabla^2\psi}{\partial z} - \frac{\partial\psi}{\partial z}\frac{\partial\nabla^2\psi}{\partial x}, \tag{4.27}$$

$$J(A,\nabla^2 A) = \frac{\partial A}{\partial x}\frac{\partial\nabla^2 A}{\partial z} - \frac{\partial A}{\partial z}\frac{\partial\nabla^2 A}{\partial x}, \tag{4.28}$$

$$J(\psi,T) = \frac{\partial\psi}{\partial x}\frac{\partial T}{\partial z} - \frac{\partial\psi}{\partial z}\frac{\partial T}{\partial x}. \tag{4.29}$$

and we have "uncurled" the induction equation. The solenoidal constraints on $\boldsymbol{B}$ and $\boldsymbol{u}$ are automatically satisfied. We adopt the stress free boundary conditions, requiring

$$\psi = \omega = 0 \quad \text{on all boundaries.} \tag{4.30}$$

The magnetic flux function must satisfy

$$A = 0, \quad \text{at } x = 0, \quad A = 1, \quad \text{at } x = L_x, \quad \frac{\partial A}{\partial z} = 0, \quad \text{at } z = 0, 1 \tag{4.31}$$

so the field remains parallel to the imposed field at the boundaries. For temperature, we impose insulated sidewalls with uniformly heated top and bottom walls

$$T = 1 \quad \text{at } z = 0, \qquad T = 0 \quad \text{at } z = 1, \tag{4.32}$$

and at the vertical boundaries we impose no heat flux

$$\frac{\partial T}{\partial x} = 0, \quad x = 0, L_x. \tag{4.33}$$

4.1.1 Linear Theory

We now discuss the main results of the linear theory of equations (4.24)-(4.26). By linearising the dimensionless system and seeking solutions of the form $\sim \exp st$ for some complex growth-rate $s \in \mathbb{C}$, we can obtain a dispersion relation

$$\beta^2(s+\beta^2)(s+Pr\beta^2)(s+\zeta\beta^2) + Pr\zeta Q\beta^2 m^2\pi^2(s+\beta^2) - RPra^2(s+\zeta\beta^2) = 0, \tag{4.34}$$

where $\beta^2 = \pi^2(1/L_x^2 + 1)$. We further introduce the normalisations

$$\tilde{s} = \beta^{-2}s, \qquad \tau = \beta^2 t, \qquad q = (\pi^2/\beta^4)Q, \qquad r = \frac{\pi^2}{L_x 2\beta^6}R \tag{4.35}$$

with which the dispersion relation can be written

$$s^3 + (1+Pr+\zeta)s^2 + [Pr(1-r+\zeta q) + \zeta(1+Pr)]s + Pr\zeta(1+q-r) = 0. \tag{4.36}$$

From equation (4.36) we can see that a bifurcation to stationary convection at Rayleigh number $r^{(e)}$ occurring when $s = 0$

$$r^{(e)} = 1 + q. \tag{4.37}$$

The normalisation introduced to the Rayleigh number rescales the Rayleigh number by the critical onset value in the field free case. If $q = 0$, we return to Boussinesq convection with no magnetic field, with $r^{(e)} = 1$ corresponding to the minimum critical Rayleigh number for the onset of convection. We can therefore see that the effect of the magnetic field is to increase the onset value of the Rayleigh number for steady convection. For overstable oscillations we require a Hopf-bifurcation meaning that $s = \pm i\omega_o$ with $\omega_0 \in \mathbb{R}$. Substitution into the dispersion relation in equation (4.36) and equating real and imaginary parts yields that overstable oscillations occur at the Rayleigh number $r^{(o)}$

$$r^{(o)} = 1 + \frac{\zeta}{Pr}(1 + Pr + \zeta) + \frac{(Pr + \zeta)\zeta q}{1 + Pr} \tag{4.38}$$

provided that the right-hand side of

$$\omega_o^2 = \frac{Pr\zeta}{1 + Pr + \zeta}(r^{(e)} - r^{(o)}) = -\zeta^2 + \frac{1 - \zeta}{1 + Pr}Pr\zeta q \tag{4.39}$$

is positive (otherwise the assumption of a Hopf-bifurcation is violated as $\omega_0 \in \mathbb{C}$ and the eigenvalue is real). An immediate consequence of equation (4.39) is that a Hopf-bifurcation always occurs at $r^{(o)} < r^{(e)}$ and hence can only precede a bifurcation to stationary convection. Further from the same equation we also require

$$\zeta < 1, \qquad q \equiv q_o > \frac{(1 + Pr)\zeta}{(1 - \zeta)Pr} \tag{4.40}$$

to ensure that $\omega_0 \in \mathbb{R}$ and we have a Hopf-bifurcation. The full picture is then as follows: if $\zeta > 1$ or $q < q_o$ then the first bifurcation to occur is the transition to stationary convection. However, if $\omega_o^2 > 0$, $\zeta < 1$ and $q > q_o$ then the first bifurcation is a Hopf-bifurcation when two eigenvalues cross the imaginary axis. As r increases, the real component of s will increase until eventually there is a bifurcation from overstable oscillations to stationary convection.

4.1.2 Derivation of a simplified nonlinear model

We now discuss the derivation of a simplified model based on physical arguements given by [7]. First let us recount that in magnetoconvection, the static state can first become unstable to overstable oscillations at $r = r^{(o)}$. As r is increased, this behaviour eventually transitions to steady convection but this does not necessarily occur at the linear theory value of $r = r^{(e)}$ but instead at $r = r^{(i)}$. Subcritical steady convection can occur at $r < r^{(e)}$ because the transition to oscillatory convection results in magnetic flux expulsion to the edge of the cells. As already noted, the magnetic field is stabilising and the expulsion of the field to the cell edges results in a field free region for which the stability properties are different. Thus depending on the values of the different parameters, convection can occur before or after $r^{(e)}$ (subcritical and supercritical respectively). For this reason a model is developed which retains higher-order terms in the Fourier series expansion [7]. Further, it is also possible for large amplitude solutions to transition from oscillatory convection to steady convection.

To derive the truncated model, we again expand our solution as a truncated Fourier series. The expansion can be written for ψ, A and T as follows

$$\psi = \epsilon 2(2p)^{1/2}\frac{L_x}{\pi}\sin\frac{\pi x}{L_x}\sin\pi z a(\tau) + \dots, \tag{4.41}$$

$$A = x + \epsilon 2(2/p)^{1/2}L_x\sin\frac{\pi x}{L_x}\cos\pi z d(\tau) + \epsilon^2\frac{L_x}{\pi}\sin\frac{2\pi x}{L_x}e(\tau) + \dots, \tag{4.42}$$

$$T = 1 - z + \epsilon 2(2/p)^{1/2}\cos\frac{\pi x}{L_x}\sin\pi z b(\tau) - \epsilon^2\frac{1}{\pi}\sin 2\pi z c(\tau) + \dots, \tag{4.43}$$

where

$$\tau = pt, \qquad p = \pi^2(1 + 1/L_x^2) = \beta^2, \tag{4.44}$$

and ϵ is a small expansion parameter. The modes chosen in the expansion are selected to be consistent with the steady weakly nonlinear problem, details of which are given in [9], though they can be understood on a physical basis. The term $\sin 2\pi x/L_x e(\tau)$ represents the concentration of magnetic flux sheets at the cell walls. The term $-\sin 2\pi z c(\tau)$ represents the formation of thermal boundary layers at the hot and cold boundaries. Substitution into the governing dimensionless equations (4.24)-(4.26) and retaining terms up to $\mathcal{O}(\epsilon^2)$ we obtain

$$\dot{a} = Pr[-a + rb + \zeta q d((\varpi - 3)e - 1))], \tag{4.45}$$

$$\dot{b} = -b + a(1 - c), \tag{4.46}$$

$$\dot{c} = \varpi(-c + ab), \tag{4.47}$$

$$\dot{d} = -\zeta d + a(1 - e), \tag{4.48}$$

$$\dot{e} = -(4 - \varpi)\zeta e + \varpi a d, \tag{4.49}$$

where

$$r = \frac{\pi^2}{L_x^2 p^3}R, \qquad q = \frac{\pi}{p^2}Q, \qquad \varpi = \frac{4\pi^2}{p}, \tag{4.50}$$

with $0 \leq \varpi \leq 4$. In future sections these will be referred to as the Knobloch model or the weakly nonlinear model. We will use equations 4.45 - 4.49 to validate SINDy behaviour. However, it is possible with differentiation of the ODEs to eliminate all variables in favour of only a (see ref. [9]) and so often we consider finite amplitude effects as being interchangable with the mode amplitude a. These equations have two distinct properties. The first is that

$$\frac{\partial\dot{a}}{\partial a} + \frac{\partial\dot{b}}{\partial b} + \frac{\partial\dot{c}}{\partial c} + \frac{\partial\dot{d}}{\partial d} + \frac{\partial\dot{e}}{\partial e} = -[Pr + (1 + \varpi) + \zeta(5 - \varpi)] < 0 \tag{4.51}$$

and so solutions may be attracted to a fixed point, limit cycle or a strange attractor. They are also invariant under the symmetry

$$a \to -a, \quad b \to -b, \quad c \to c, \quad d \to -d, \quad e \to e. \tag{4.52}$$

The advantage of this system of ODEs is that it facilitates a much faster investigation of the results from both linear and weakly nonlinear theory. It can also characterise the behaviour of sub and supercritical convection which cannot be understood from linear theory alone. The results of linear theory can be found by neglecting all quadratic terms and searching for solutions of the form $\sim \exp(s\tau)$. In this case we obtain the dispersion relation

$$s^3 + (1 + Pr + \zeta)s^2 + [Pr(1 - r + \zeta q) + \zeta(1 + Pr)]s + Pr\zeta(1 - r + q) = 0 \tag{4.53}$$

which is exactly (by construction) the dispersion relation already given. So all previous statements on the solution behaviour apply in the linear case. Similarly we may consider the finite amplitude solutions about the bifurcation at $r^{(e)}$ when $\omega_o^2 < 0$ by expanding

$$r = r^{(e)} + r_2^{(e)} a^2 + \mathcal{O}(a^4) \tag{4.54}$$

where the expansion is chosen to preserve the symmetry $a \to -a$. By equating powers of a, we find

$$r_2^{(e)} = 1 + q + \frac{(2-\varpi)\varpi q}{(4-\varpi)\zeta^2} \tag{4.55}$$

which is the result expected from weakly nonlinear theory. This correction tells us when to expect subcritical and supercritical behaviour. If ζ is small, then $r_2^{(e)}$ is positive if $\varpi < 2$ implying tall cells and negative for flat cells. We can also conclude that as r increases through $r^{(e)}$, finite amplitude solutions will be unstable (subcritical) if $r_2^{(e)} < 0$ and stable (supercritical) if $r_2^{(e)} > 0$ provided $\omega_o^2 < 0$. In total these equations represent a simplified model of the full-order system which explain the effects of finite amplitude perturbations. The equations can successfully described the bifurcation structure of the full PDE system [7]. This model will be used to assess the performance of SINDy as it displays a wide variety of behaviours. We will mainly be interested in cases where some type of oscillation is present, like with overstability. Instabilities in tokamaks typically exhibit oscillations in the diagnostics, hence why we primarily limit our interest to this case and not steady convection (see ref. [16]).

4.2 SINDy applied to the Knobloch equations

We now assess the performance of SINDy on equations (4.45)-(4.49) to understand where limitations of the method may appear. In general there are several limitations that are given by SINDy [97] including: optimisation challenge due to increasing library size, quality of derivative estimates from noisy observations, quality of derivative estimates in poorly sampled signals and correlated features in the library. We will first discuss some model evaluation metrics which are chosen for their speed of evaluation. We will then apply SINDy to some parameter cases discussed in ref [9] and the selection of learning parameters for different optimisers. We will also apply SINDy across solutions from the oscillatory branch and in general note the sensitivity of SINDy to errors introduced from finite differencing. As a result we consider the application of weak SINDy and discuss the selection of optimisation hyper-parameters. We also build a series of constraints for SINDy in relation to properties of the outlined system of equations which help promote physically sensible models in noisy, poorly sampled cases.

4.2.1 Model evaluation metrics

To evaluate model performance we consider two different performance metrics. The first is the mean-squared error

$$\text{MSQE} = \frac{1}{nm} \sum_{i=0}^{i=n-1} \sum_{j=0}^{j=m-1} (\dot{X}_{ji} - \hat{\dot{X}}_{ji})^2, \tag{4.56}$$

where $\dot{X}$ are the true derivatives and $\hat{\dot{X}}$ the predicted derivatives from SINDy. The result is averaged over all n equations. We calculate the mean-squared error for both the training data and a reserved set of validation data, though we calculate these on noiseless data. It is also common to calculate a mean absolute coefficient error on the non-zero coefficients in the identified model [152]

$$\text{MACE} = \frac{1}{nk_{nonzero}} \sum_{j=0}^{n-1} \sum_{i=0}^{i=k_{nonzero}-1} \frac{|C_i - \hat{C}_i|}{|C_i|} \tag{4.57}$$

where C_i is the array of true coefficients. The result is again averaged over all equations. In future this will be referenced to as the coefficient error. In all examples discussed below we limit our analysis to a second-order polynomial library formed from all input time series, unless explicitly stated otherwise. The library then consists of a constant term, linear terms in a,b,c,d and e as well as all quadratic products of the linear terms without repetition.

4.2.2 Validation of learning parameters for oscillatory convection

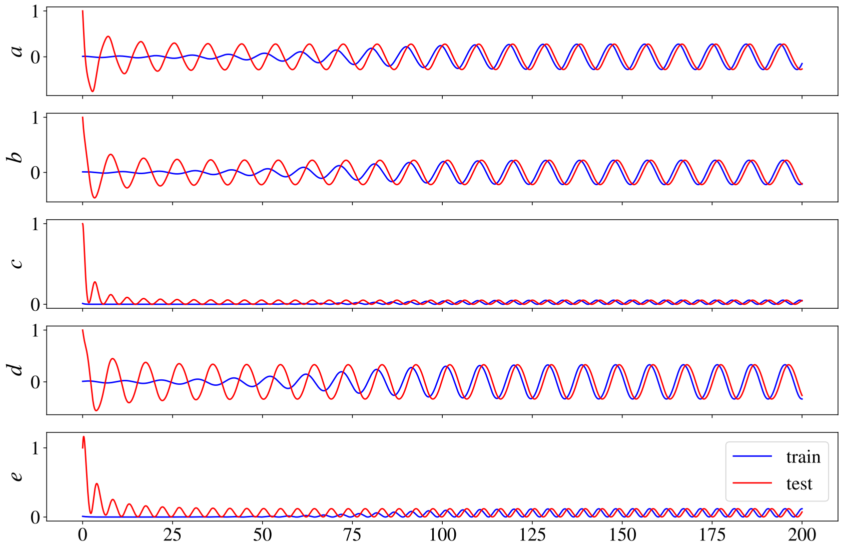


Figure 33: Train (blue) and test (red) data for the parameters $Pr = 1$, $r = 3.6$, $q = 5$, $\zeta = 0.4$ and $\varpi = 2$. An initial condition of 0.001 for all variables is used in the training data, and 1 for all variables in the test data.

For the sequentially-thresholded least squares optimiser, there are two learning parameters which must be fixed. These are the coefficient threshold λ which controls the maximum allowable coefficient magnitude and the L2 regularisation α which also controls the sparsity of the final solution (larger α promoting sparser solutions). To choose these parameters we test SINDy on a fixed set of data taken at the parameter values $\zeta = 0.4, Pr = 1, r = 3.6, q = 5, \varpi = 2$ with equations (4.45)-(4.49) with small initial conditions shown in Figure 33, along with the corresponding test trajectory. The choice of parameters is close to the transition from stable to oscillatory convection and produces overstable solutions. Both λ and α are varied on a logarithmically spaced scale between 0.0001 and 1 in 40 steps. At each combination of λ and α the errors are calculated. The training data is generated at a time-step of $\mathrm{d}t = 0.001$ and then downsampled to a sampling rate of $\nu = 200$. Limiting the sampling rate is important, as it impacts the quality of the derivative estimates.

Figure 34 shows the calculation of the three error metrics on a grid with varying α and λ as well as the total number of non-zero coefficients found at each α and λ. In each figure we also plot a corresponding colourbar which indicates the value of the error or the total number of coefficients in the case of the bottom right subplot only. Each plot then represents an error landscape, where the coefficient threshold increases in magnitude in the positive x direction and the L2 regularisation increases in the positive y direction with errors calculated at each combination of α and λ. For identified SINDy models, model selection follows the principle of Occam's razor: the simplest model to describe the data is best. This means that as λ is increased, we expect to obtain progressively sparser models as we progressively remove features which do not describe the and therefore these models will all obtain low errors. As λ is increased we expect to reach a threshold beyond which a key feature is removed and the error suddenly rises. The region just preceding this point represents the sparsest handful of models which best describe the data.

Starting from the top left of Figure 34, the train MSQE is plotted, and shows a large region of low error indicated by light blue given that $\lambda \lesssim 0.2$ which is insensitive to the choice of α. Regions of low error indicate that SINDy has identified models which faithfully represent the training data. This behaviour is reasonable as the error increase occurs when λ becomes larger than the smallest expected coefficient in the underlying ODE system. Comparison of the train MSQE to the test MSQE shows that not all low error train MSQE regions correspond to low errors region with the test MSQE. In this case it occurs if either the coefficient threshold λ or the L2 regularisation becomes too large. The fact that we encounter low error training regions and high error testing regions indicates that SINDy has identified a model which is not generalisable and is not capable of reproducing the dynamics from the reserved data-set.

Comparison of the train MSQE to the coefficient error in Figure 34 again shows that just because a model obtains a low train MSQE, it does not imply we recover the true underlying system. Again areas which obtain a low train MSQE have a high coefficient error meaning we do not recover the true underlying system. If we look at the number of non-zero coefficients obtained, we can see that all models which have a low train MSQE also have the densest models (highest number of active terms). Inspection of these models indicates that the correct active terms are identified along with many small non-zero coefficient terms (much smaller than the threshold value). While this seems to disagree with the STLSQ process, the routine used in pysindy differs from the method implemented in ref [97]. The final step involves an “unbiasing” step where an unregularised least-squares fit is performed on the remaining features which can result in coefficients smaller than the chosen threshold. The unbiasing approach is different to the approach originally used in [97] and may produce models with small non-zero terms. A fit to the data with $\alpha = 0$ confirms that this is the case, as the resulting model has the expected form and the expected number of non-zero coefficients, with no coefficient values smaller than the threshold. Overall the performance of SINDy is not clear in this simple test-case. Despite the data being noiseless and there existing a known true sparse model, SINDy instead recovers the correct system with many accompanying small non-zero terms.

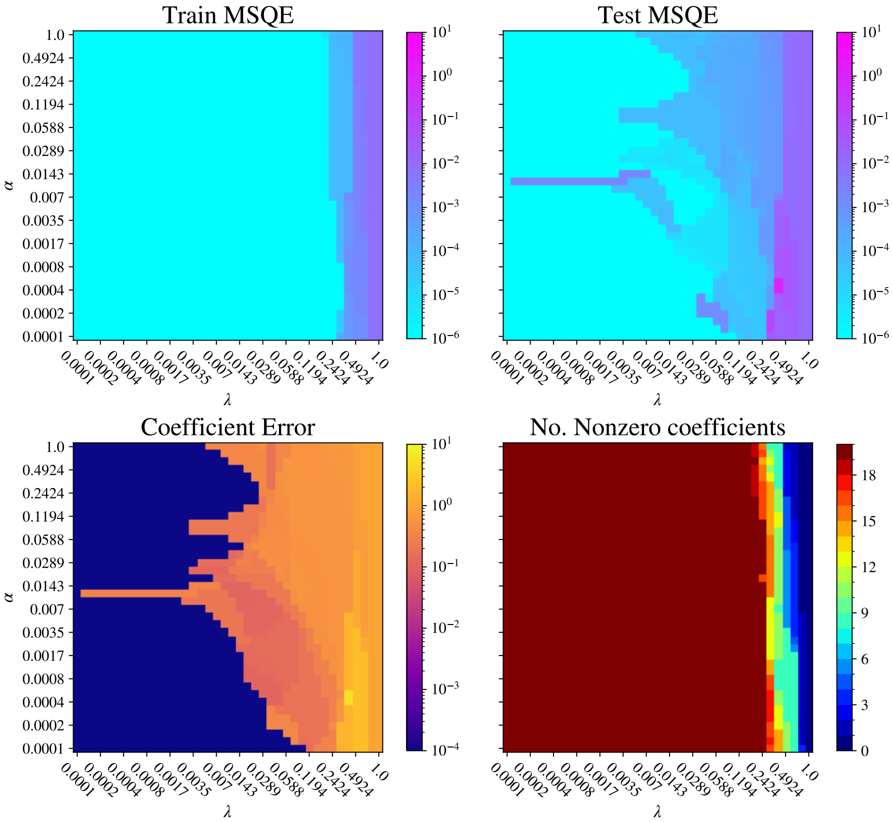


Figure 34: Plots of calculated errors while varying the coefficient threshold λ and L2 regularisation α for training data close to the onset of overstable oscillations. Low error regions in the train MSQE show models fitting the training data well, but these do not necessarily obtain low MSQE on test sets or recover the true underlying models.

4.2.3 Variation of learning parameters with sampling rates

One aspect which is often not discussed in SINDy literature is the dependency of the learning parameters on the sampling rate and errors within the data. Two cases are considered: the first when the derivatives are estimated from a fine time-step of $\mathrm{d}t = 0.001$ using finite differencing and subsequently downsampled to a sampling rate of $\nu \approx 50$ after they have been calculated (we term this pre-computed derivatives). The second is when the sampling rate is limited to $\nu \approx 50$ and the derivatives are estimated using finite differencing on the already downsampled training data. In the first case we therefore expected fewer identification errors related to errors from estimation in the derivatives. We then compare the optimal learning parameters identified in each case.

Figure 35 shows the results for pre-computed derivatives at a sampling rate of $\nu = 50$ where the errors have been calculated on a grid of varying α and λ. Despite the derivatives being calculated on highly sampled data, the resulting boundary between the low coefficient error region and high coefficient error region differs from that shown in Figure 34. The only difference in training data in this comparison comes from the sampling rate used in identification. This creates a challenge, as usually hyper-parameters are taken at the elbow of the error curve, where we aim for the resulting model to be as sparse as possible but still attain a low error (corresponding to large λ). However, selection of optimal parameters in one case does not necessarily carry over to another case in which the sampling rate of the data is slightly lower, even though the same accuracy of derivatives is used in both cases.

Results where derivatives are computed from the downsampled data are shown in Figure 36. The train MSQE still shows a similar region of successful identification to Figure 35, but the models now produce much higher errors in both the test MSQE and coefficient error. Such a result shows that SINDy is capable of easily identifying models which reproduce the training data, but are not generalisable or good estimates of the true underlying system. We can see that the errors are reduced closest to the Pareto frontier but these models still have many non-zero coefficient terms. This simple example demonstrates the challenge SINDy faces on noiseless data even when sampling rates are arguably high. The difficulty arises due to correlation in the feature library and input time series and in this case SINDy cannot identify sparse or generalisable models, even when sparser answers exist. While the correlation is more obvious here, it will certainly be less obvious in more general examples but can cause a significant deterioration of the SINDy model identification process.

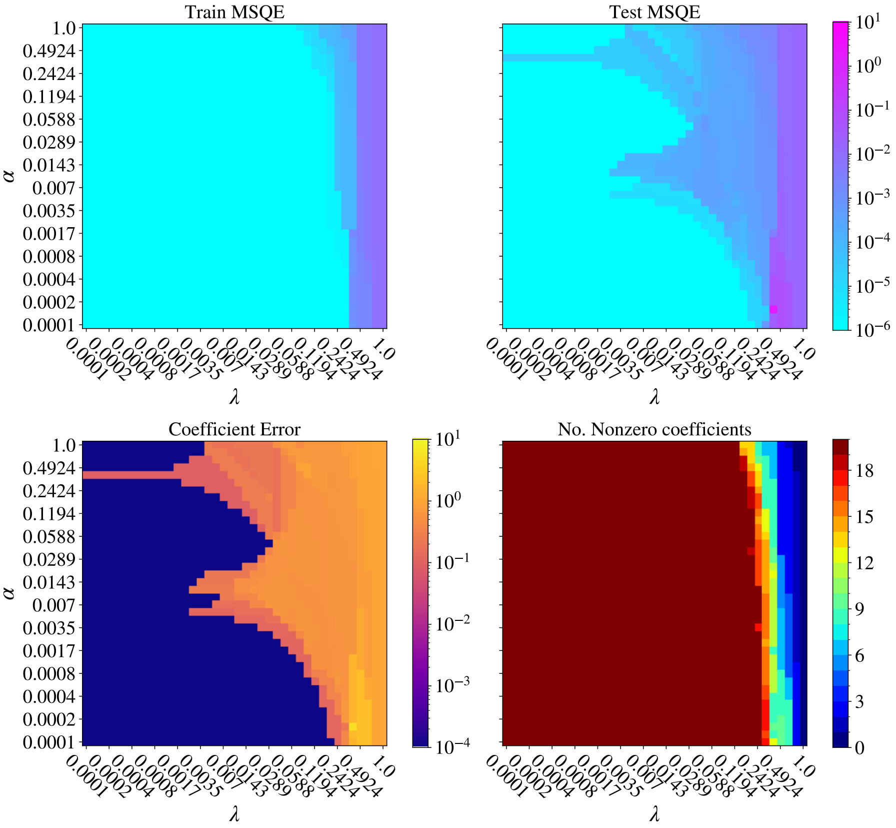


Figure 35: Calculated errors for varying L2 regularisation α and coefficient threshold λ where the derivatives are estimated from finely sampled data at $dt = 0.001$ and subsequently downsampled to $\nu = 50$.

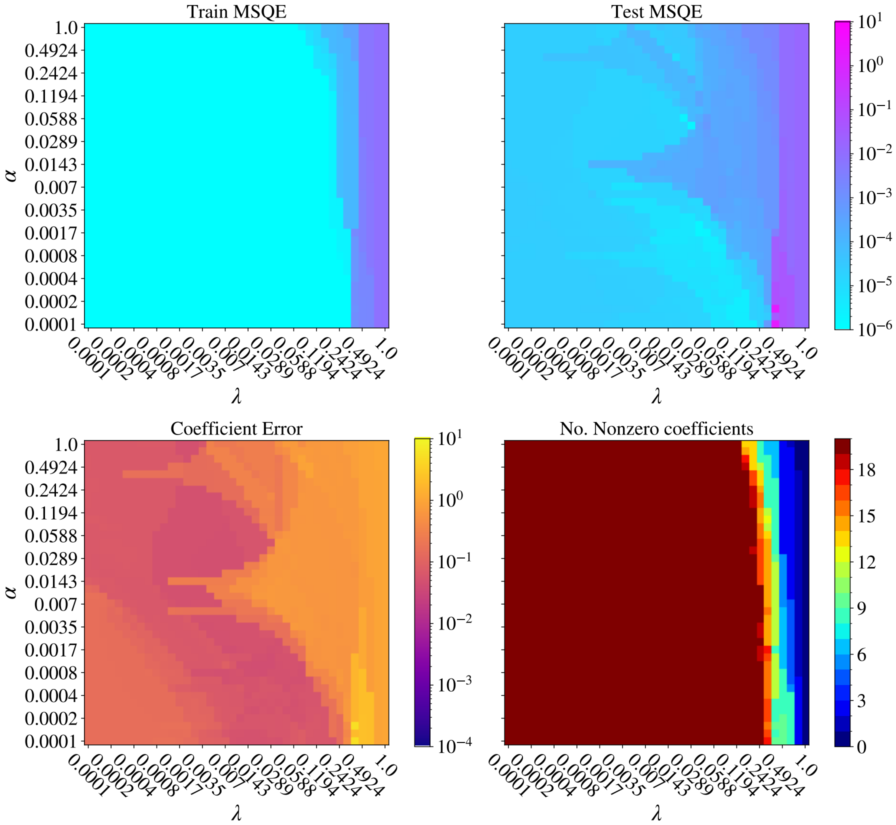


Figure 36: Calculated errors for varying L2 regularisation α and coefficient threshold λ, where the derivatives are estimated using finite differencing with data already downsampled to $\nu = 50$.

4.2.4 Noise sensitivity

We now assess the performance of SINDy when noise is added to the data. We first consider the case where additive noise is added to the states and the derivatives are calculated on noiseless data. We should expect this approach to yield more robust results for two reasons. The first is that finite differencing amplifies the noise present in the signal, so taking derivatives of noisy series will yield increased noise levels. The second is that the library terms are nonlinear, and when noise is added to the library terms, noise can be further amplified when products of the feature terms are taken. By adding noise to the derivatives after they have been calculated, we avoid both of these issues. For this test case we add a very small amount of noise with standard deviation $\delta_\sigma = 0.001$ and limit the sampling rate to $\nu \approx 50$.

The results are shown in Figure 37, where the train MSQE has been calculated on the noiseless data and not the noisy derivatives. We can see that despite the noise being extremely minor in this case, there is a substantial increase in the test error and coefficient error. This is largely due to the identification of many non-zero terms which, due to the presence of a small amount of noise, no longer have small coefficients following the unbiasing step. For example the model with the minimum test MSQE has the form

$$\dot{a} = -1.004a + 3.610b - 2.004d - 0.008ae + 0.030be - 2.017de, \tag{4.58}$$

$$\dot{b} = 1.007a - 1.017b + 0.007d - 0.990ac - 0.004ae - 0.015bc + 0.008de, \tag{4.59}$$

$$\dot{c} = -1.980c + -0.013e + 0.004a^2 + 1.983ab + 0.010d^2, \tag{4.60}$$

$$\dot{d} = 1.006a - 0.015b - 0.394d - 1.006ae + 0.016be - 0.008cd, \tag{4.61}$$

$$\dot{e} = -0.794e - 0.062a^2 + 0.277ab + 1.877ad - 0.307b^2 + 0.275bd - 0.067d^2 - 0.013e^2 \tag{4.62}$$

which, alongside the correct feature terms, identifies many small coefficient terms. Some coefficients are below the threshold value due to the aforementioned unbiasing step. In the overstable case, STLSQ fails to correctly identify the correct modes due to correlation in the input time-series. Given the sensitivity of this case to noise, even denoising methods would provide little benefit. It is also interesting to note that despite sparser models describing the data being available (e.g., the true one), SINDy does not select these models. Further, while increasing α should promote sparser solutions, it appears that the additional penalisation causes poor recovery of the true terms. The case of overstable oscillation therefore seems to present a challenging case for SINDy despite the apparent simplicity.

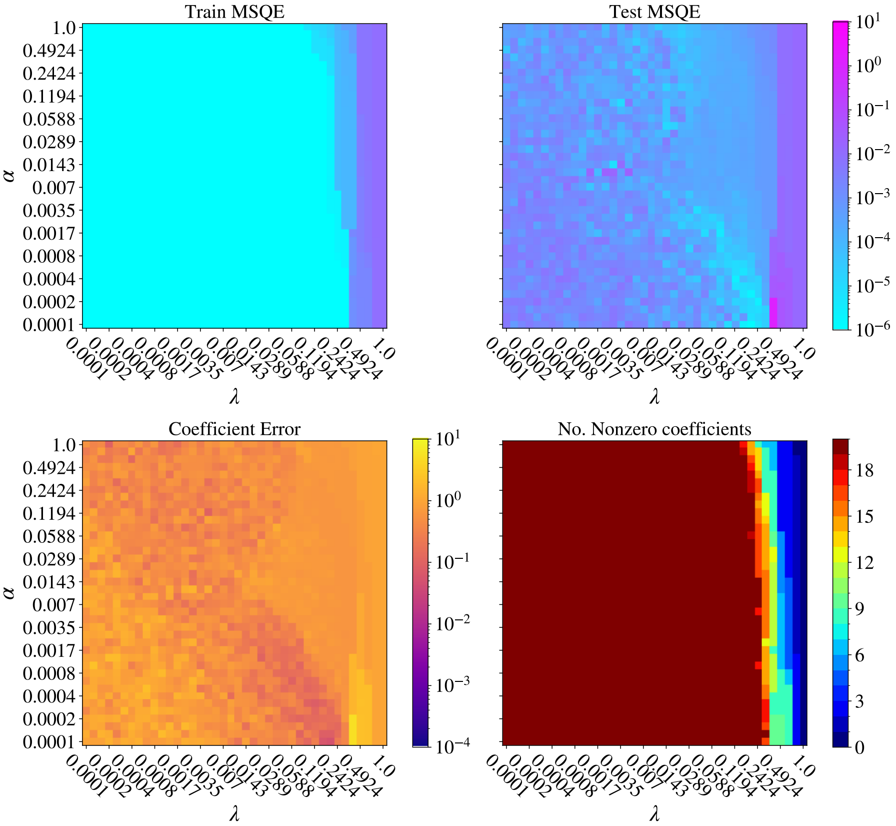


Figure 37: Plots of different error metrics while varying the coefficient threshold λ and L2 regularisation α with noise added to the derivatives with standard deviation $\delta_\sigma = 0.001$.

4.2.5 Using large initial conditions to improve model identification

One of the primary issues with robust model identification in the overstable case is using initial conditions which do not show different mode evolutions. In the case of starting from a small amplitude perturbation, SINDy struggles to separate the correlated features in the library. One solution is to start instead at the initial condition given by $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$ where the amplitude of the nonlinearities is much larger. We first repeat the previous case with noise added to the derivatives with value $\delta_\sigma = 0.001$ and $\nu \approx 50$. The results are shown in Figure 38 where we now see a distinct improvement in all errors. In particular we see that the region corresponding to the lowest coefficient errors now finds models which have the desired number of non-zero coefficients.

We also show the case with noise added with standard deviation $\delta_\sigma = 0.1$ in Figure 39. With this new initial condition, we can see that the results are substantially more robust to additive noise which agrees with the assumed form of noise added to the linear regression (only noise on the calculated derivatives). In this case there is still a band of α and λ which gives models with low coefficient errors and the correct number of non-zero coefficients. We also see that in this instance, the regions of low train MSQE and test MSQE agree, as opposed to identifying models which have low training errors but incorrect active nonlinearities.

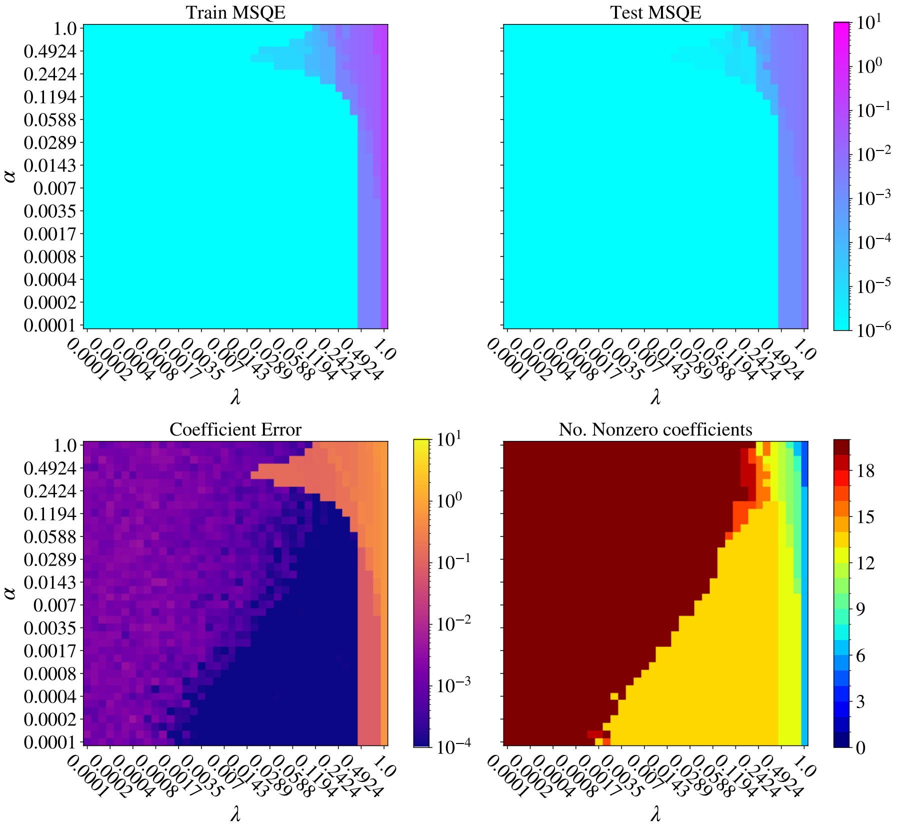


Figure 38: Evaluation of different errors with noise while varying learning parameters when SINDy is trained on a trajectory with initial condition $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$ with additive noise on the derivatives with value $\delta_\sigma = 0.001$.

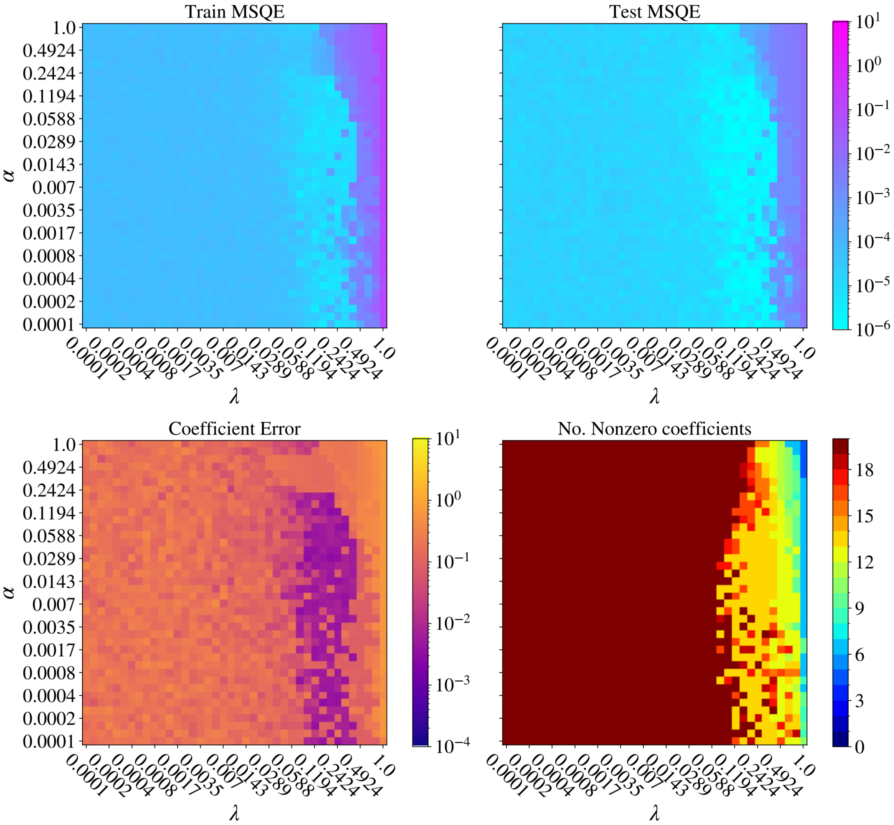


Figure 39: Evaluation of different errors with noise while varying learning parameters when SINDy is trained on a trajectory with initial condition $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$ with additive noise on the derivatives with value $\delta_\sigma = 0.1$.

4.2.6 Noise robustness with measurements of state

The more relevant case of noise robustness are cases where noise is added to state measurements. The derivatives are calculated using finite differences which results in the noise being amplified in the calculation of the derivatives. The issues are two-fold. The first results from the amplification of the noise due to the finite differencing scheme, the second comes from the composition of the feature library. Linear terms results in additive Gaussian noise to the library features, but nonlinear product terms of the state also amplify the noise and can create non-Gaussian noise distributions clearly violating the Gaussian noise assumption in linear regression.

We start by considering random Gaussian noise added to the states with standard deviations $\delta_\sigma = 0.001$ and $\delta_\sigma = 0.01$. For $\delta_\sigma = 0.001$ shown in Figure 40, model identification remains good with both low test and coefficient errors. At higher noise levels shown for $\delta_\sigma = 0.01$ in Figure 41, we can see that model identification is substantially poorer with low coefficient error cases only appearing on occasion. However, on inspection the low coefficient error solutions contain many incorrect non-zero terms. For example, one equation for $\dot{d}$ identified with minimum coefficient error has the form

$$\dot{d} = 2.283b - 1.260d - 1.001ae + 2.251b^2 - 0.673bc - 2.602bd - 4.698be - 0.574c^2 + 5.086cd + 0.627ce + 0.629d^2, \tag{4.63}$$

which has fair estimates for the expected terms, but contains many non-zero incorrect terms. While the coefficient error is reasonably small, many incorrect terms are also identified. This is because the coefficient error is relative scoring method, and is not calculated on terms which should be zero. As the calculation involves computation of the relative coefficient error, if the coefficient should be zero then the score would not be defined.

If the models with minimum errors are integrated against a validation set, their performance is poor suggesting that this degree of noise is already too substantial for SINDy to cope with, with the resulting models being overfit to noise. The issue can be understood simply from Figure 42. We can see that $\delta_\sigma = 0.01$ added to the state constitutes a small amount of noise for this system, however, when the finite differences are computed the derivatives are extremely noisy. Even for small degrees of noise signal with $\delta_\sigma = 0.001$, the derivatives are again very noisy.

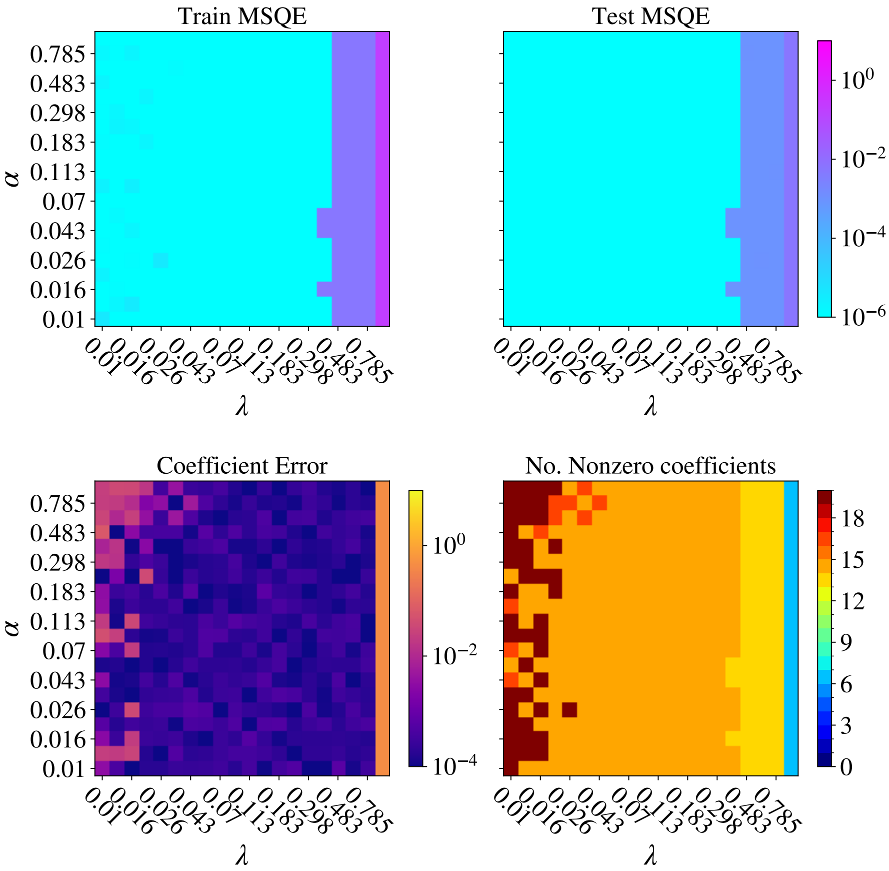


Figure 40: Comparison of the errors with noise added to the state trajectories (rather than directly to the derivatives) with standard deviation $\delta_\sigma = 0.001$ with initial condition $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$. In each grid the L2 regularisation α and coefficient threshold λ are varied.

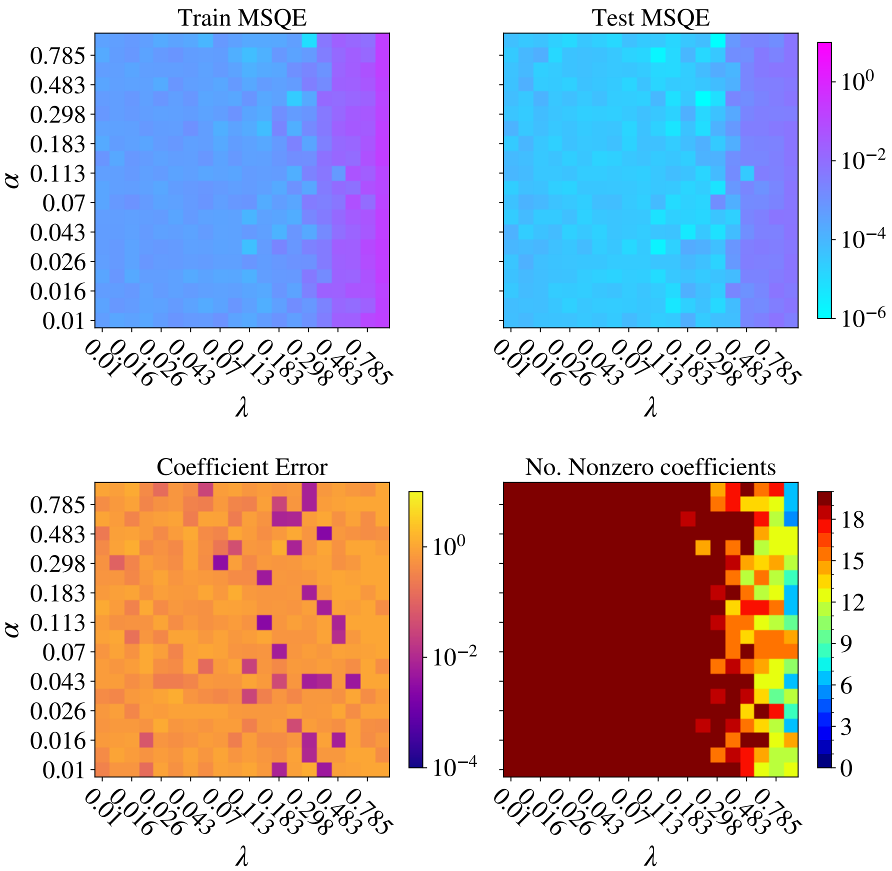


Figure 41: Comparison of the errors with a larger degree of noise added to the states with standard deviation $\delta_\sigma = 0.01$ and initial condition $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$. In this case noise is amplified by using finite differencing to estimate the derivatives. In each grid the L2 regularisation α and coefficient threshold λ are varied.

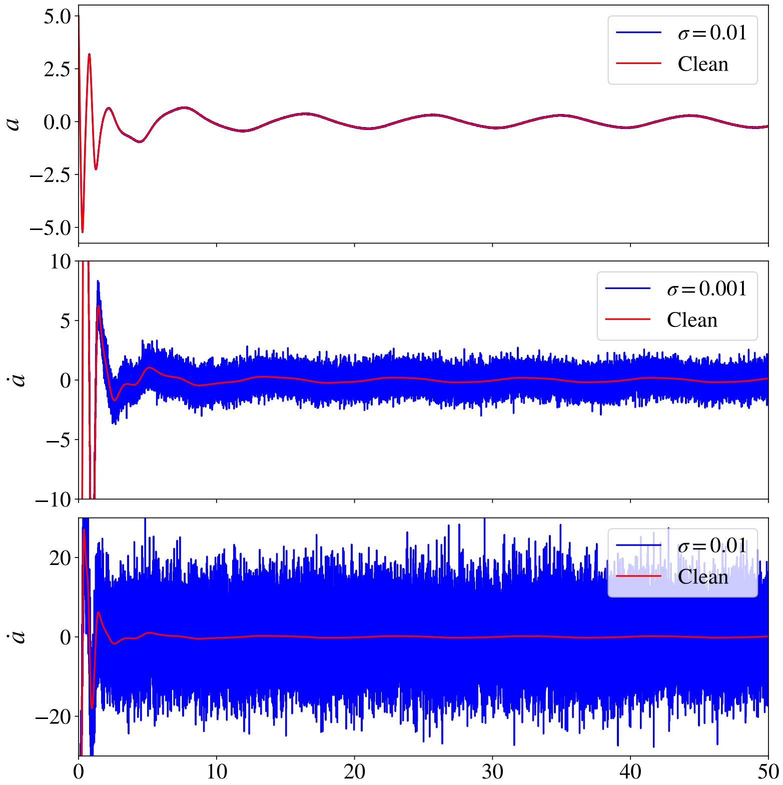


Figure 42: A comparison of the true train data in red and the noisy state data for a in blue (top). The next two columns compare the computed derivatives in blue for $\delta_\sigma = 0.001$ (middle) and $\delta_\sigma = 0.01$ (bottom) against the true derivatives in red.

4.2.7 Restriction of the modes

For parameters close to the onset of oscillatory convection with initial conditions close to $\mathbf{0}$, SINDy does not achieve robust identification of the underlying model. Minor degrees of additive noise or even lower sampling rates can cause poor identification of the true model. As many of the input time series are correlated, we can try simply leaving some of the correlated time series out of the regression. This is also an interesting test-case for SINDy as it typically requires full-state measurements. Given this, we restrict a SINDy fit to either the components related to ψ and T (a, b and c), or the components related to ψ and A (a, d and e). In the following regressions we fix the learning parameters to $\lambda = 0.1$ and $\alpha = 0.01$ and $\nu = 50$. We first begin by fitting a model with only 3 modes. Fitting a model to a, d and e results in

$$\dot{a} = 0.465a - 0.569d - 1.209ae - 0.526de, \tag{4.64}$$

$$\dot{d} = 1.000a - 0.400d - 1.000ae, \tag{4.65}$$

$$\dot{e} = -0.800e + 2.000ad \tag{4.66}$$

which identifies the expected equations for d and e but in a the buoyancy term represented by b has been replaced by a quadratic term in ae, though all coefficients are different from their expected values which should be

$$\dot{a} = -a + 3.6b - 2d - 2de. \tag{4.67}$$

A comparison of the model performance to the test and training data is shown in Figure 43, where we can see that the data is well reproduced except in the case where the initial condition is small and the initial growth rates are not correctly estimated. Similarly we can see at larger amplitudes of initial conditions, the nonlinearities will have larger amplitudes and this highlights that the identified model is only generally capable of fitting on attractor dynamics. This model does not appear to be generalisable to different initial conditions at fixed parameter values.

We also compare this three mode fit to a fit of four modes of a, b, c and d shown in Figure 44. In this case starting from small initial conditions still produces models which are out of phase with the true underlying one, but still saturate at the correct amplitudes. The results here suggest that we can at least, for fixed parameter values, fit SINDy models which are incomplete in state-space measurements. However there are some important considerations. For example, the model in a, d and e idenfities no dependence on the temperature for the $\dot{a}$ equation, and specifically no linear b term corresponding to the buoyancy. While SINDy aims to provide sparse and generalisable model, in this case we can identify a model which does not show the influence of temperature gradient on convection.

Further, it is also not clear how many modes and which combination of modes should be used. One notion is to simply exclude the modes with the smallest amplitudes. Results for a fit of a, b, d and e shown in Figure 45 appear to show a better fit than the 4 mode fit given by a, b, c and d. If a fit is performed with the initial condition $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$ then we do not get sparse models like those shown above. Instead it is apparent that the restricted selection of modes cannot capture the initial large amplitude transients and hence the resulting models are non-sparse.

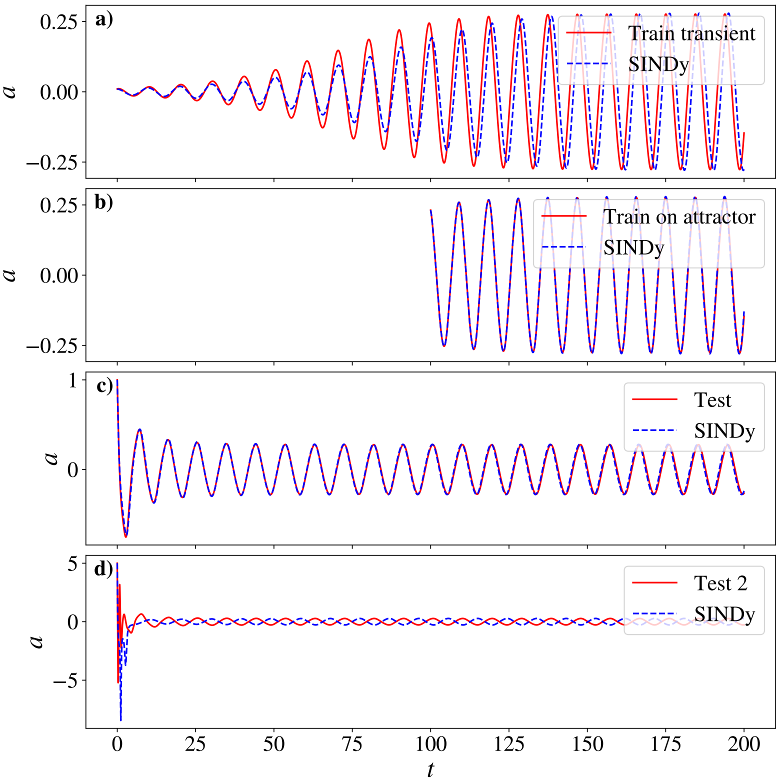


Figure 43: A comparison of the performance of the identified SINDy model fit to a, d and e on the a) full training data, b) only a section on the stable attractor in the training set, c) test set with initial condition $\boldsymbol{a}_0 = (1, 1, 1, 1, 1)$ and d) initial condition with $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$.

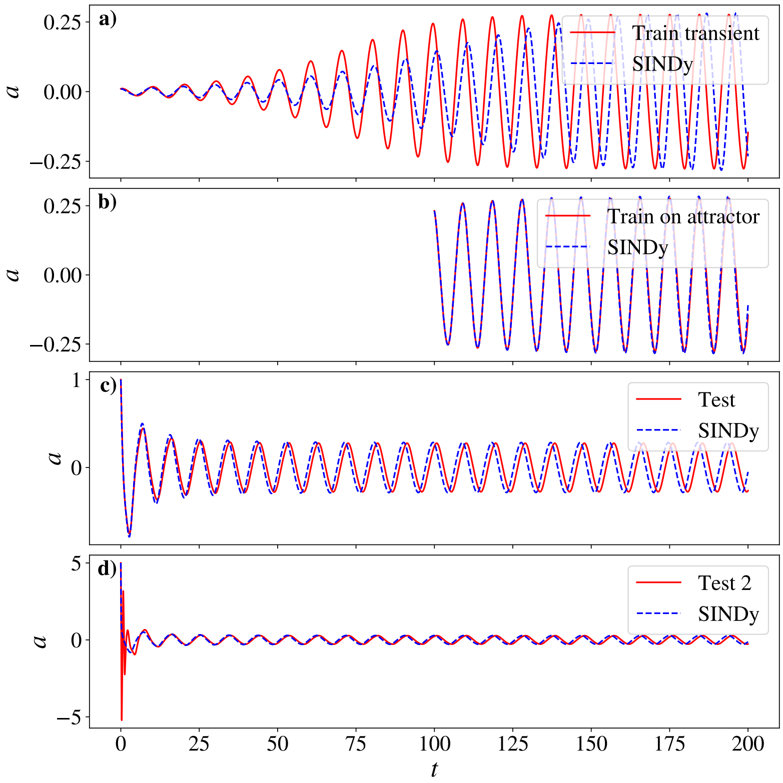


Figure 44: A comparison of the performance of the identified SINDy model fit to a, b, c, d on the a) full training data, b) only a section on the stable attractor in the training set, c) test set with initial condition $\boldsymbol{a}_0 = (1, 1, 1, 1, 1)$ and d) initial condition with $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$.

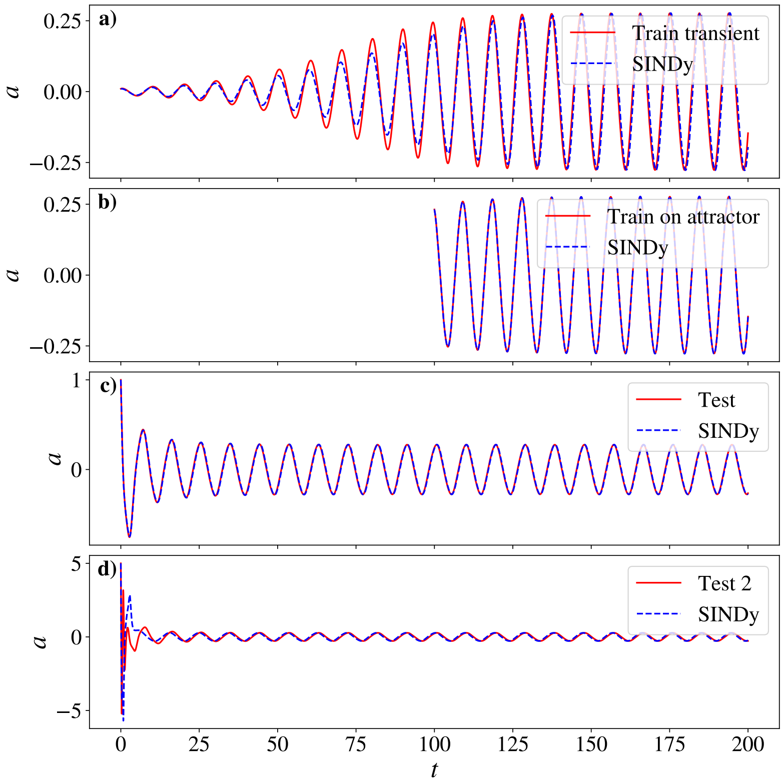


Figure 45: A comparison of the performance of the identified SINDy model fit to a, b, d, e on the a) full training data, b) only a section on the stable attractor in the training set, c) test set with initial condition $\boldsymbol{a}_0 = (1, 1, 1, 1, 1)$ and d) initial condition with $\boldsymbol{a}_0 = (5, 5, 5, 5, 5)$.

4.2.8 Impacts of nonlinearity

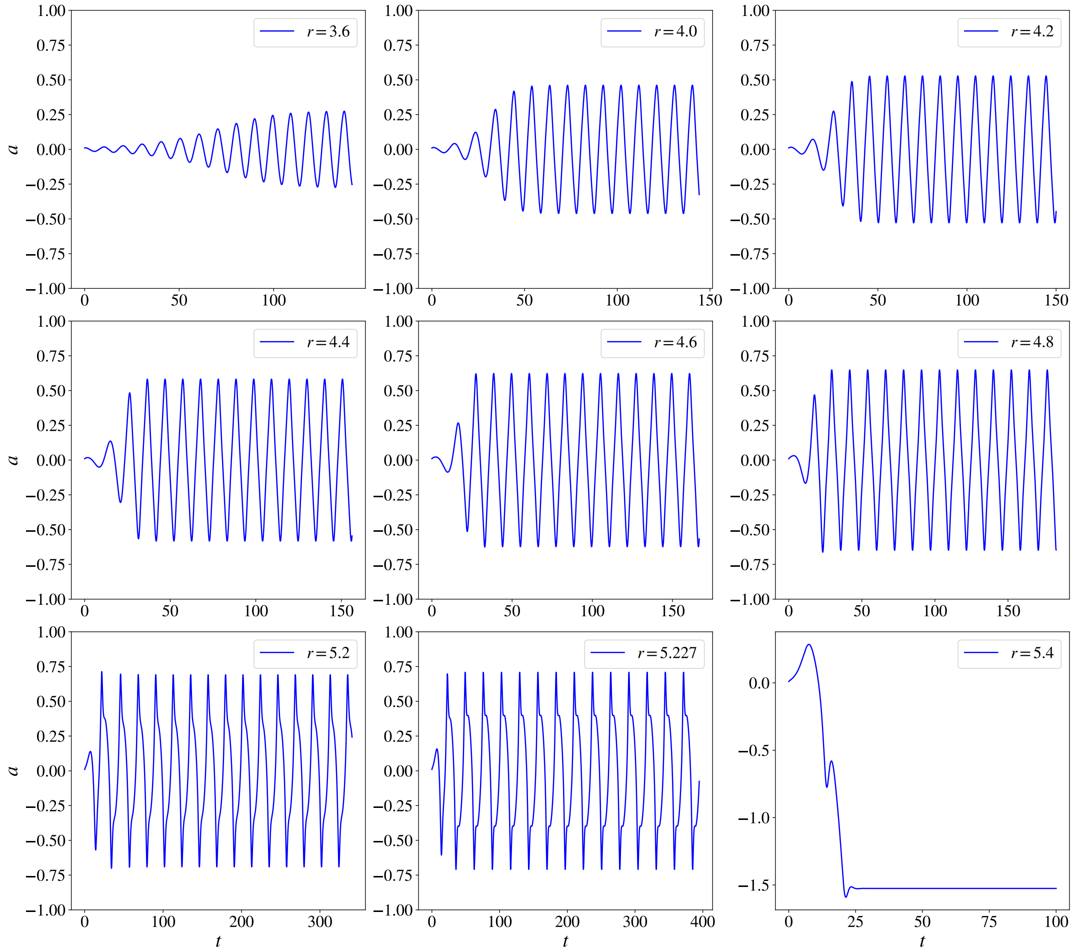


Figure 46: Illustration of the training data showing the series generated from a as the normalised Rayleigh number r is varied. As r increases the oscillations become progressively nonlinear until there is a transition to steady convection.

As we increase r along the oscillatory branch, eventually we transition from overstable oscillations to steady convection. As we approach the bifurcation, the oscillations become notably more nonlinear and the period increases, shown in Figure 46. We assess the performance of SINDy in these conditions while keeping the number of samples per period and the total number of periods fixed. Training trajectories are generated by varying r and fixing the other parameters to $q = 5$, $Pr = 1$, $\zeta = 0.4$, $\tilde{\omega} = 2$. We set $\nu_{max} \approx 200$ and the total number of periods $N_L = 30$. For the steady case, it is no longer possible to define periods of the system and instead we set the maximum time to $T = 500$. For each value of r up to the bifurcation, the periods are explicitly found as $[9.43, 9.62, 10.0, 10.42, 11.11, 12.20, 22.73, 26.36]$, in agreement with those listed in ref [13].

Figure 47 shows the calculated coefficient errors when the derivatives are calculated by finite difference and Figure 48 shows the coefficient errors for derivatives pre-calculated with a time-step of $\mathrm{d}t = 0.001$. For $r < 4.8$, the test error remains relatively consistent with the regions of λ and α producing low scoring models staying the same. However, at $r = 5.2$ both the coefficient error and test error increase substantially. Closer to the bifurcation at $r = 5.227$, model identification is even poorer with no learning parameters appearing to give successful identification even at relatively high sampling rates. Results from pre-calculated derivatives in Figure 48 show that these errors are entirely to do with the accuracy in the derivative estimation. As we approach the bifurcation point, the period increases but this in turn gives poor estimates of the derivatives due to the sharpness of the oscillations. Aside from errors arising in derivative estimations, the increasing non-linearity of the signal appears to have no impact on the selection of valid hyper-parameters.

Figure 49 shows the coefficient error calculated while varying the total number of periods of data and sampling rate, with pre-computed derivatives. In this instance $\lambda = 0.1$ and $\alpha = 0.0017$ as suitable learning parameters from the α, λ parameter sweeps. Contrary to the previous examples, increasing r appears to allow successful equation identification at lower sampling rates provided enough data is supplied. Further, as r increases it is possible to have successful model identification with fewer total periods of data, provided the sampling rate is high enough. This is because closer to the onset of oscillatory convection, the transient growth is slower and it takes longer to reach the attractor. For $r = 4.2$, this happens after 5 periods for the highest sampling rates which is still before attractor has been reached, shown in Figure 46. However, in all cases it at least one period of data is needed for these initial conditions.

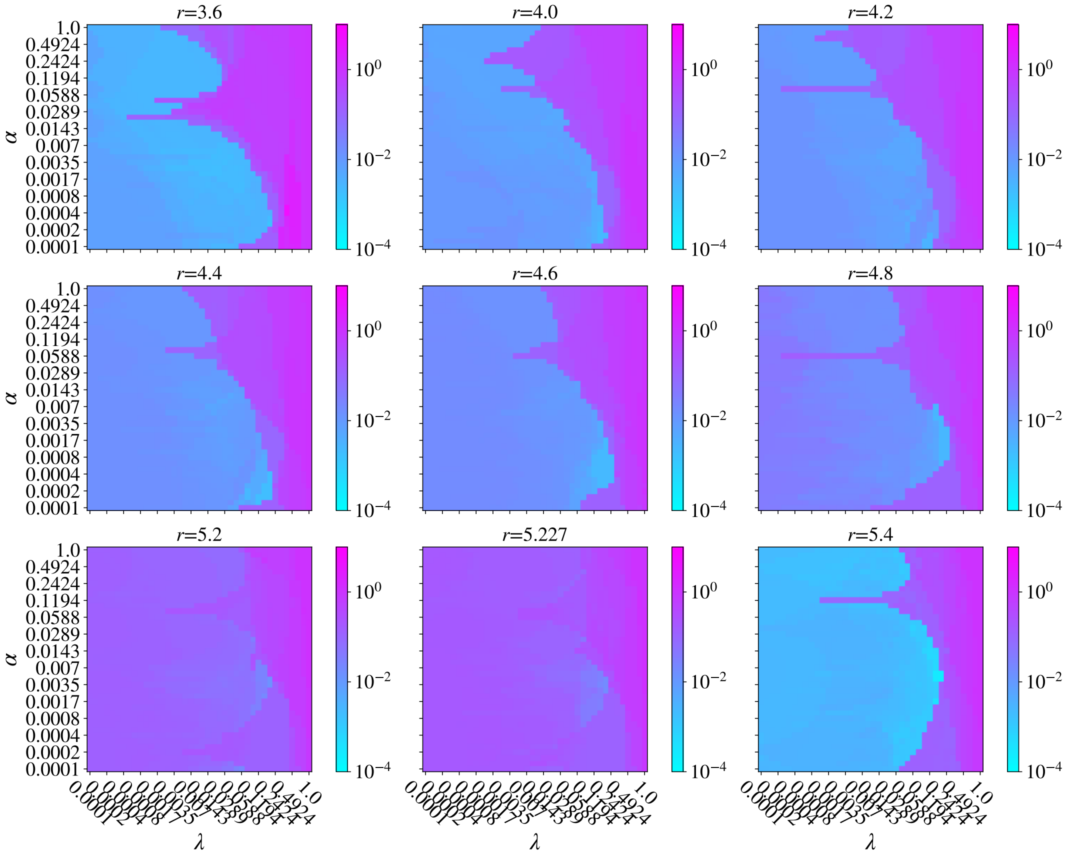


Figure 47: Coefficient errors for varying the L2 regularisation α and the coefficient threshold λ with fixed sampling rate $\nu = 200$ where derivatives are calculated from finite differences. Each box shows the errors calculated as the training data changes with varying r.

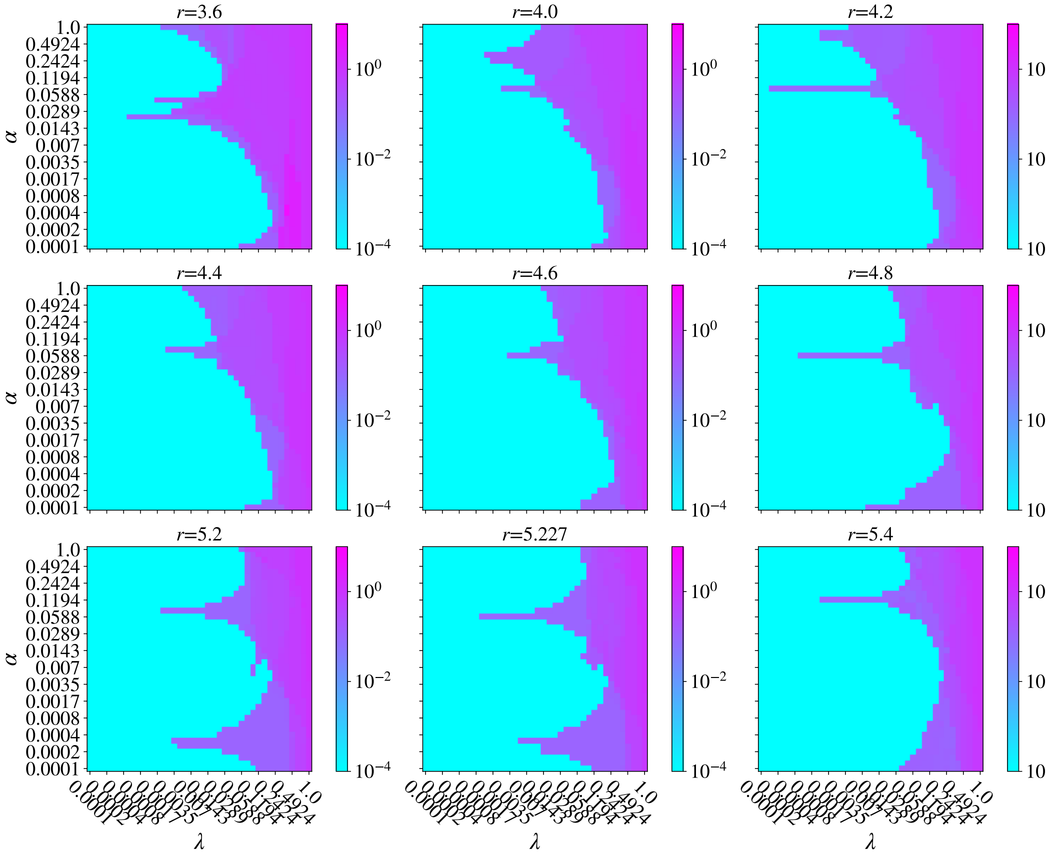


Figure 48: Coefficient errors for varying α and λ with fixed sampling rate $\nu = 200$ where derivatives are pre-calculated from a data-set with $t = 0.001$. Each box shows the errors calculated along the oscillatory branch as r varies, highlighting the impact of nonlinear oscillations in the identification process.

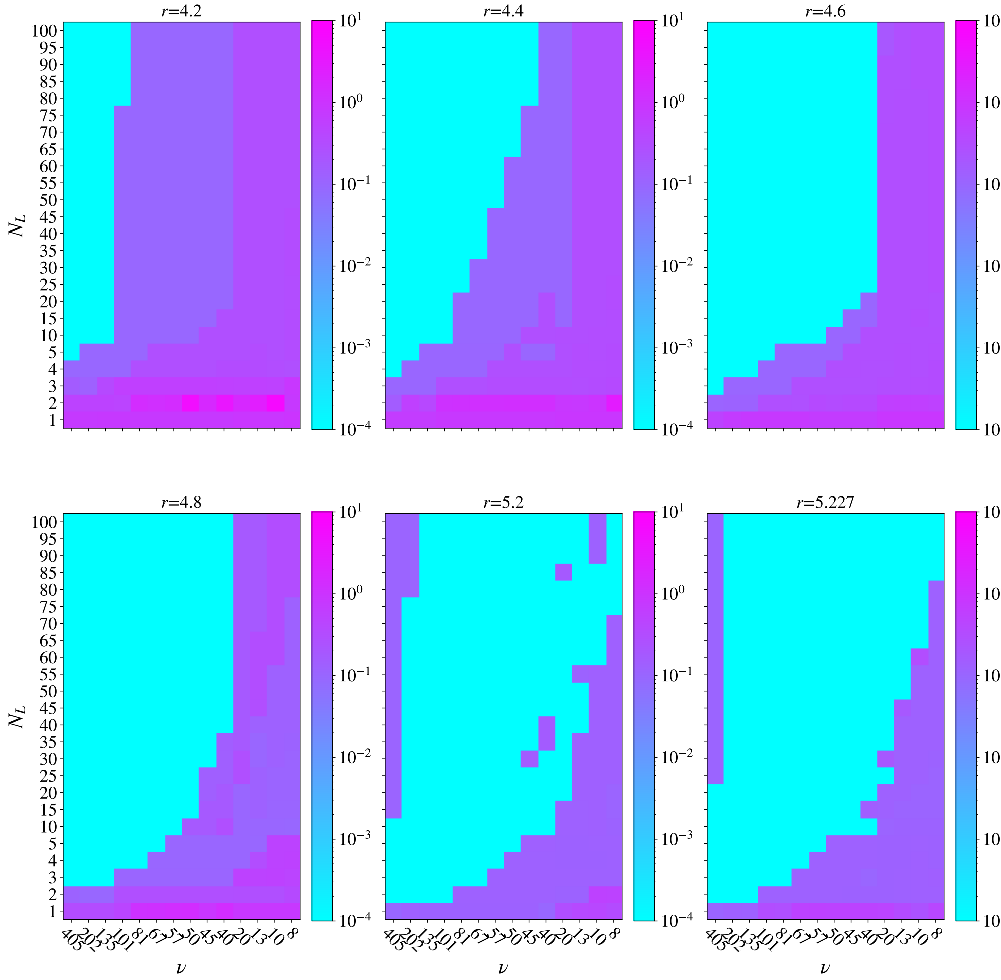


Figure 49: Sampling requirements with fixed learning parameters at coefficient threshold $\lambda = 0.1$ and L2 regularisation $\alpha = 0.0017$ as the maximum number of periods (N_L) and the sampling rate (ν) vary. Errors calculated on a selection of normalised Rayleigh numbers is shown.

4.2.9 Period doubling identification

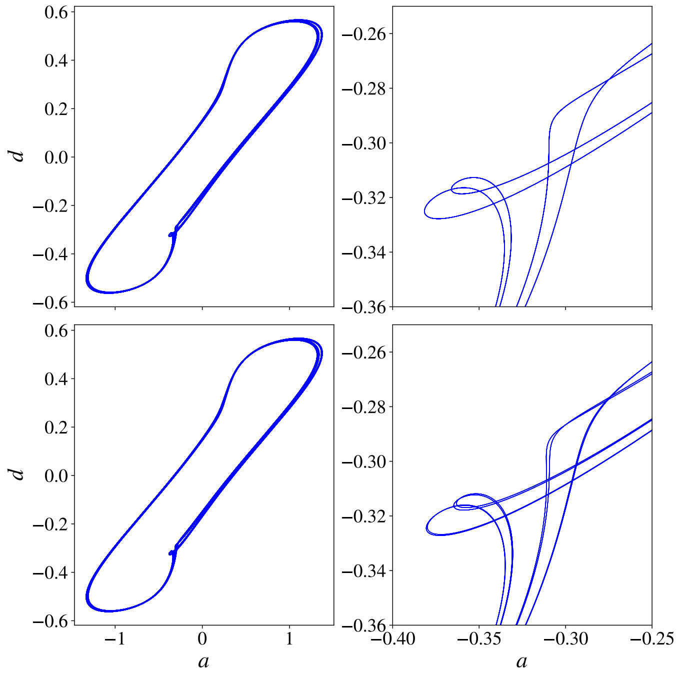


Figure 50: Example of period doubling with period 4 in the top row with $r = 5.29$, $Pr = 9.5$ and period 8 in the bottom row with $r = 5.291$, $Pr = 9.5$.

For different parameter values, the ODE system exhibits period doubling followed by a transition to chaos. Similar to studying the impact of nonlinearity on successful equation recovery, here we see if trajectories which take a longer time to return to their starting location in phase space are more or less challenging for SINDy. For the period doubling parameters considered, all cases have the same dominant period given by the FFT with period $L = 13.45$. While variation of the parameters does make the period double, this does not change the trajectory significantly enough to alter the dominant frequency, with this time corresponding to approximately one revolution in the $a - d$ plane, illustrated in Figure 50. To test the sensitivity of SINDy to period doubling, we compute the coefficient error for varying sampling rates and total numbers of periods of the data.

The coefficient error is shown in Figure 51 for the select parameters which result in period doubling. We can see that there appears to be no benefit to studying either longer period signals or semiperiodic signals. The error plots in these cases are near-identical despite trajectories taking a longer time to return to their starting location in the phase-space. When the derivatives are pre-computed, there are instances where we achieve successful equation recovery when $\nu \approx 13$ if the data is long enough.

The results are now shown when the derivatives are computed from finite differencing in Figure 52. We can see that quite a different picture is obtained in this case, and the correct model is only obtained for the highest sampling rates and with a certain length. At these parameter values the fastest time-scales of the system require much higher sampling rates to be properly resolved.

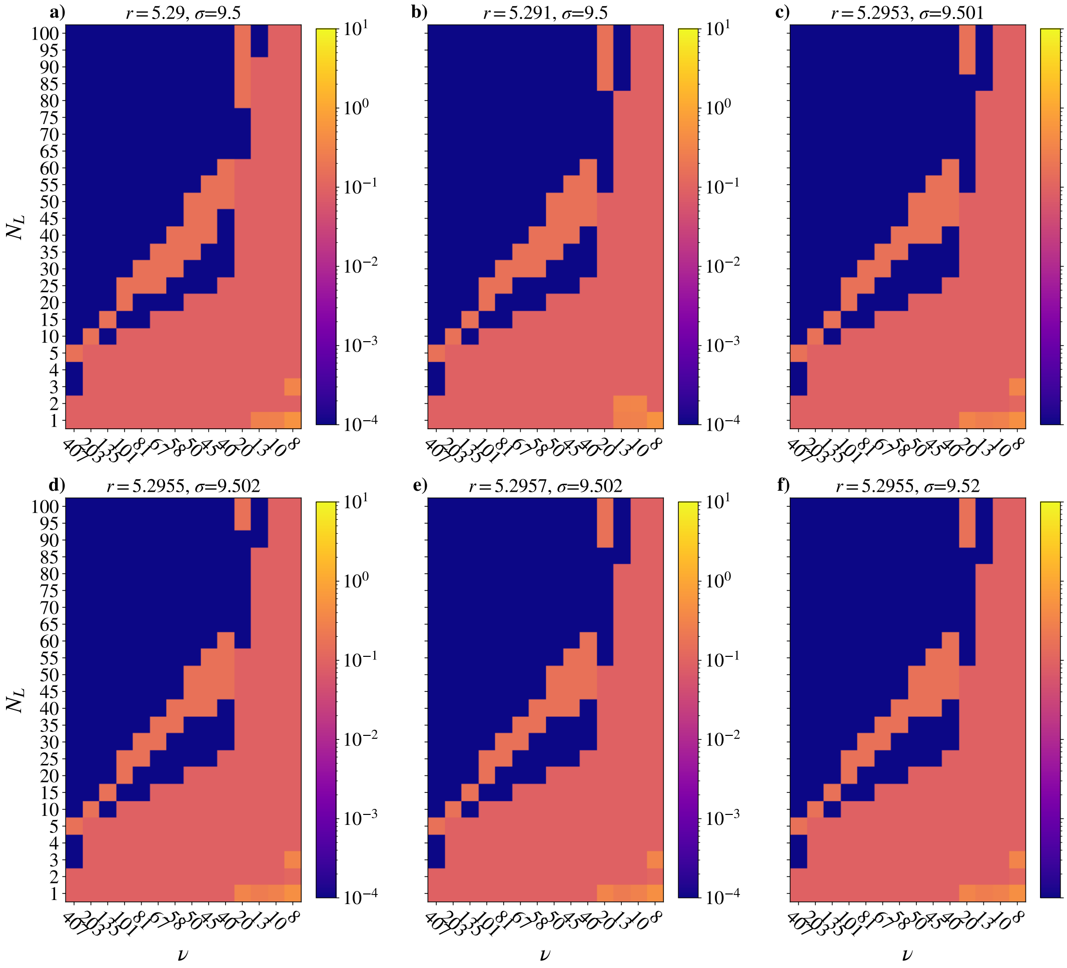


Figure 51: Coefficient error for varying r and Pr which exhibits period doubling as the values are changed. The derivatives are pre-calculated on $dt = 0.001$. a) period 4, b) period 8 c) period 12, d) period 24 e) period 48 and f) semi-periodic. Here the number signifies the number of revolutions in the ad plane before returning to its original starting location.

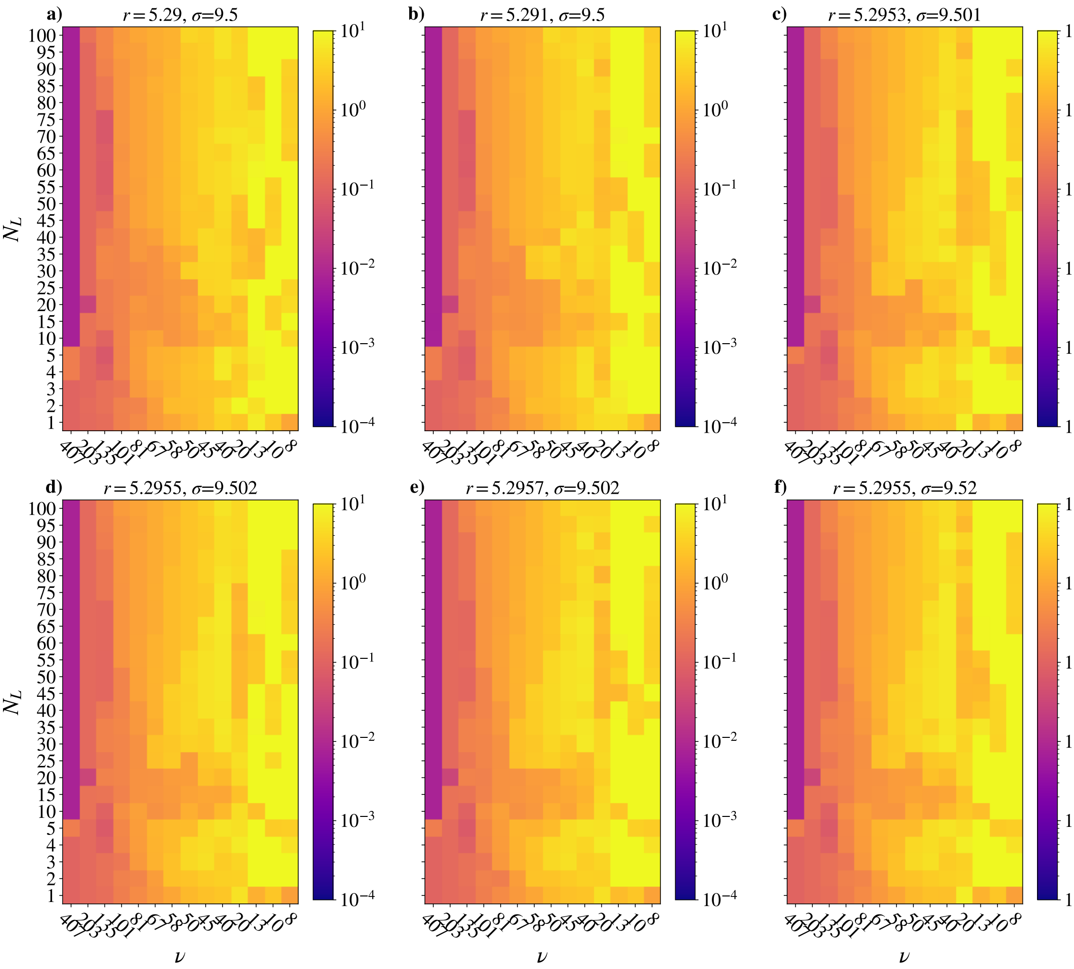


Figure 52: Coefficient error for varying r and Pr which exhibits period doubling as the values are changed. The derivatives are pre-calculated on $dt = 0.001$. a) period 4, b) period 8 c) period 12, d) period 24 e) period 48 and f) semi-periodic. Here the number signifies the number of revolutions in the ad plane before returning to its original starting location.

4.2.10 Weak SINDy

Two primary issues identified so far are the strong dependence on high sampling rates for SINDy and a sensitivity to noise. Most of the results from studying this system support the conclusion that the sensitivity to both sampling rates and noise arise from the computation of the derivatives. Even in cases with hundreds of samples per period, we have identified several noiseless cases where model identification is poor. One resolution to both of these issues is to use the weak form of SINDy which relies on an integral formulation of the library features and objective. With the weak form, provided there are no implicit terms, the derivatives can be completely transferred from the library terms to test functions with well defined derivatives given by the coefficients of fitted polynomials.

In the weak form, there are several additional free parameters that must be chosen. The most important of these is the integration domain window sized term H_{xt} which fixes the size of windows over which integrals are performed. This must be chosen in such a fashion that noise is suitably averaged over, but also not too large such that periodic signals are averaged over. Most cases we consider have some degree of periodicity and so windows must be typically taken to be less than one period of the characteristic time-scale of the system. We perform an assessment of the coefficient error for a fixed sampling rate of $\nu = 50$ per dominant frequency for $Pr = 9.5$, $r = 5.29$,

$\varpi = 2$, $\zeta = 0.4$ and $q = 5$ while varying the integral window size and the learning parameters α. In this example, the number of integration domains (K) is fixed to a large value at $K = 1000$ where any increase does not change the nature of the results.

The results are shown in Figure 53. It appears that larger α causes a greater importance on smaller integration domain sizes. However, in this case a smaller α creates a more robust regression problem producing better results for a much wider choice of H_{xt}. Unsurprisingly there is an upper limit to α beyond which model identification is poor for any selection of H_{xt} meaning that the solution has too strong a preference to minimisation of the L2 norm. The selection of H_{xt} appears to have a strong dependence on the periodicity of the signal. Windows lengths approaching 1 period of the system begin to produce poor results averaging out the signal itself. Window lengths that are too short contain insufficient information on the signal and thus produce poorer estimates of the coefficients.

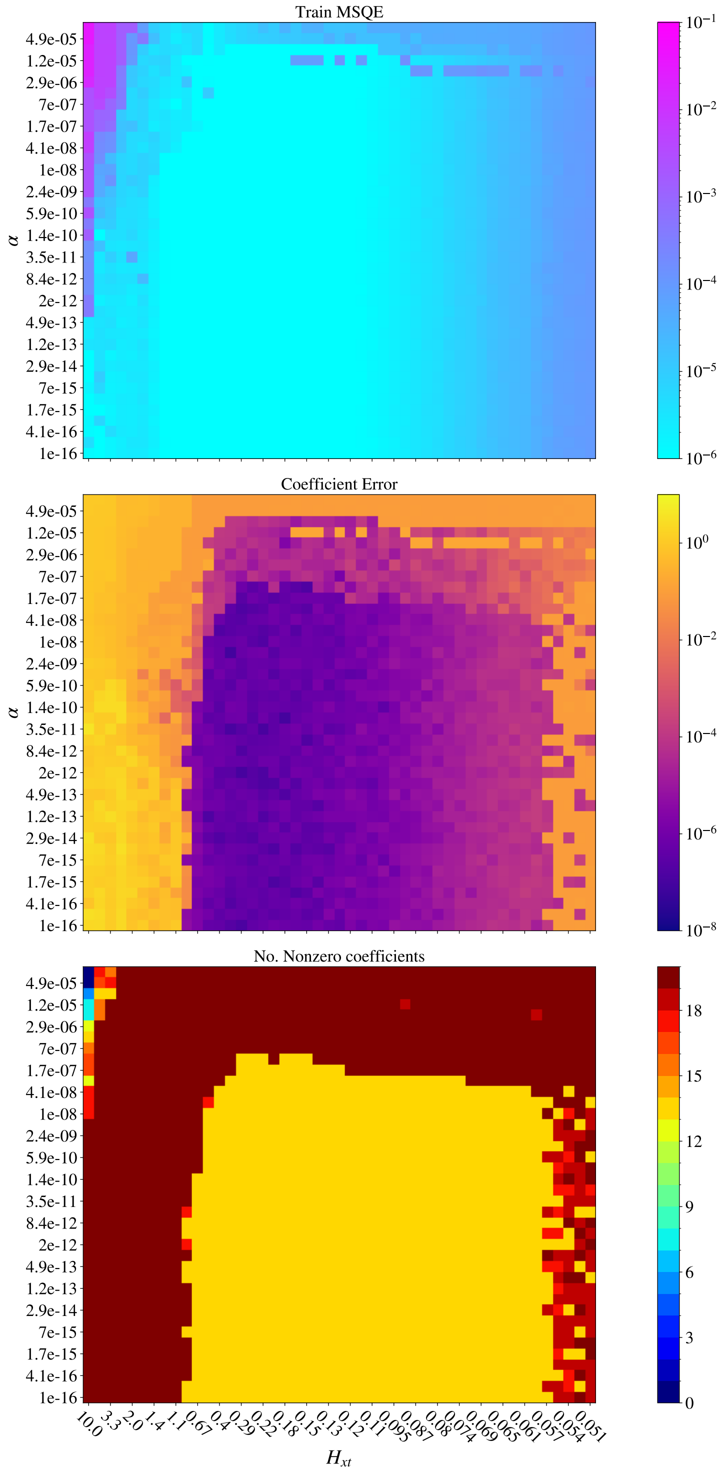


Figure 53: Different calculated errors for the weak formulation while using STLSQ while varying the L2 regularisation α and the integration window size H_{xt}. The training data is taken at a fixed instance of the period doubling parameters.

We next evaluate the model identification of the weak form by varying the total number of periods of data and the sampling rate, keeping $H_{xt} = 0.1$ period and $\alpha = 1 \times 10^{-16}$ fixed. From the result of errors in Figure 54 we can see that the coefficient error remains low for $\nu \geq 40$. As ν is increased from $\nu = 40$, the coefficient error decreases suggesting a decreasing error from better approximation of the integrals in the weak form. We note that the weak form is much more robust at lower sampling rates even though no derivatives are provided here. Even in the best case scenario with provided derivatives in Figure 51 sampling rates could rarely be lower than $\nu = 20$ and suggest the performance of the weak form is nearly equivalent. For lower sampling rates, the weak form successfully identifies the model for much shorter lengths of data than conventional SINDy with pre-computed derivatives. The lack of strong dependence on the length of the training trajectory is not surprising as the number of integral domain windows and the size of these windows does not depend on the overall length of the time series. Even over one period there are still $K = 1000$ formed integrals and in the noiseless case no averaging over noise is required. Still, this is remarkable as in the case with $\nu \approx 50$ and one period, the weak form successfully identifies the correct model. For $K = 1000$ integration domains and $H_{xt} = 0.1$ we will almost certainty have overlapping integration domains and therefore many samples will contain no new information.

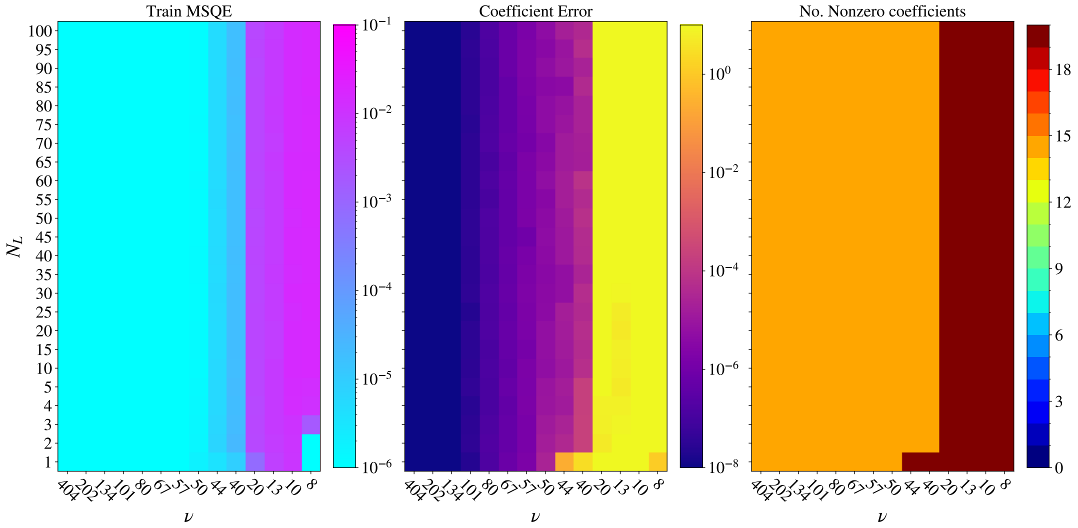


Figure 54: Calculation of different errors with fixed $L2$ regularisation α and integration window size H_{xt} while varying the numbers of periods of training data N_L and the sampling rate ν. The Figure studies the requirements of total data and fidelity of data when using the weak form.

4.2.11 Comparison to the Lorenz system and assessment of noise robustness

A useful case to consider is when the Knobloch system produces chaotic Lorenz type solutions when the magnetic field is decoupled from the fluid flow ($q = 0$), for which the remaining parameters are set to $r = 28$, $\varpi = 8/3$, $Pr = 10$ and $\zeta = 0.2$. In this case we can compare fits between the first three equations (a, b, c) which yields the Lorenz system and a higher-dimensional problem for a, b, c, d, e. We can then compare how increasing the dimension of the optimisation problem impacts the noise robustness in this case. Increasing the number of equations increases the library size, but can also introduce correlated features. In this way increasing the number of equations may present a harder optimisation problem than simply increasing the size of the polynomial library for a fixed number of equations.

To classify the time-scale in this case, instead of using the dominant time scale from an FFT we instead use the inverse of the maximal positive Lyapunov exponent ≈ 0.906 as a characteristic time-scale [43][pg. 431]. We further use a fine time-step at $dt = 0.001$ for initial assessment of the integration windows and limit the total length of data to 100 Lyapunov times.

We first discuss the results of fits to a, b, c, d and e with 1% noise added to the training data, shown in Figure 55. We can see that there is a wide range of H_{xt} and α which produce low coefficient errors. Window sizes exceeding one Lyapunov times produce poor estimations of the underlying model. Window sizes that are substantially smaller than 0.1 Lyapunov times give correct active terms in the underlying model (not shown) but produce poorer estimates of the coefficients. This is likely because the noise is not averaged over. For 10% noise shown in Figure 56 equation recovery is significantly poorer, though low coefficients regions still persist. In constrast, when only a, b and c are fit shown in Figure 56 for 10% noise, the correct equation is still recovered for a band of H_{xt}.

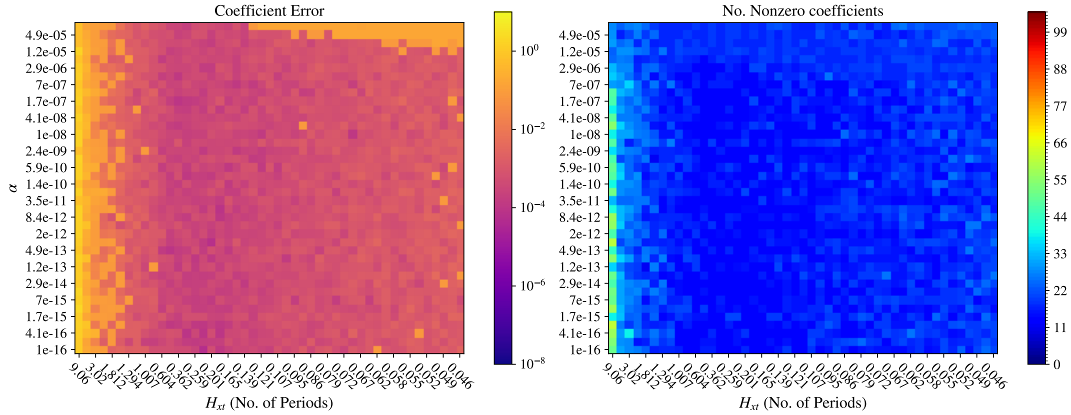


Figure 55: Variation of the L2 regularisation α and the size of the integration domain H_{xt} using the weak form for a complete set of input time series a, b, c, d, e and noise with standard deviation 1% of each variable.

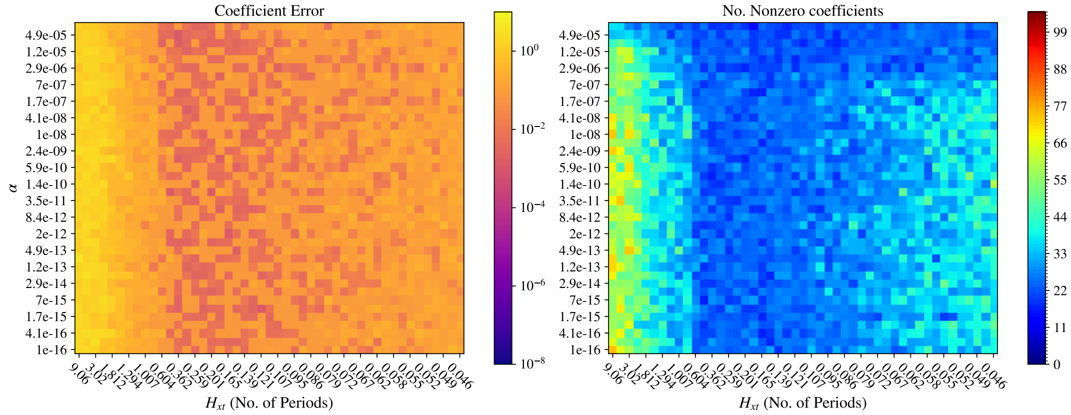


Figure 56: Variation of the L2 regularisation α and the size of the integration domain H_{xt} using the weak form for a complete set of input time series a, b, c, d, e and noise with standard deviation 10% of each variable.

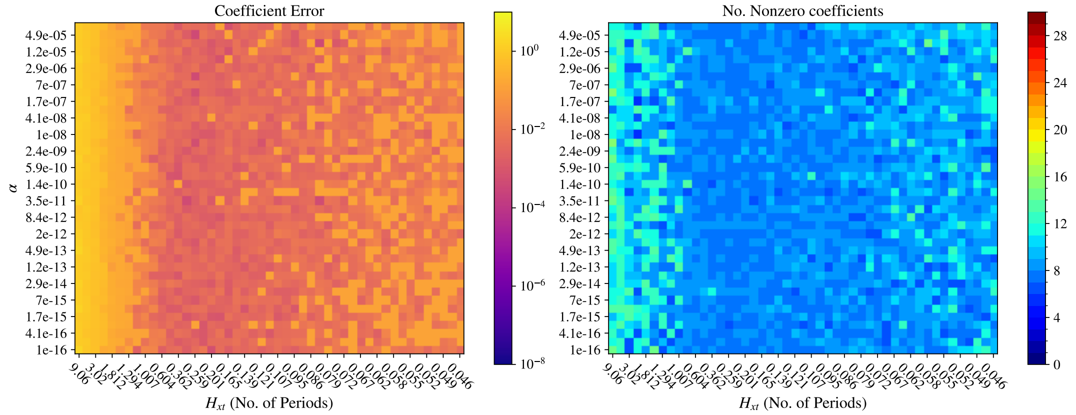


Figure 57: Variation of the L2 regularisation α and the size of the integration domain H_{xt} using the weak form using only the Lorenz time series a, b, c as input with noise with standard deviation 10% of each variable.

Again we repeat the calculations of errors for the cases described above varying the sampling rate ν and the total number of periods N_L over which the entire regression is performed for $K = 1000$ windows. For this regression we fix H_{xt} as 0.906/3 determined previously and $\alpha = 1 \times 10^{-8}$. The results are first shown for a, b, c, d, e for 1% noise in Figure 58. In this case weak SINDy provides the correct model reliably at sampling rates as low as 100 samples per Lyapunov time though cases with lower sampling rates still produce good estimates of the true coefficients but introduces additional non-zero terms which are not in the true model. For lower sampling rates, the equations are typically poorly recovered regardless of the length of data provided. The fact that error does not always decrease with increasing N_L for low sampling rates implies that the particular selection of integration domains are not linearly independent at a given seed. In other words, for lower sampling rates the data must be uniquely sampled with non-overlapping integration domains. Unsurprisingly the coefficient error reduces with increasing sampling rate. More interestingly, there appears to be a minimum length of data at around 15 Lyapunov times beyond which increasing the sampling rate does not reduce this requirement. This is related to the training data itself, where the first 15 Lyapunov times are located on one lobe of the Lorenz attractor only.

We next compare 10% noise in Figure 59 for all 5 equations and Figure 60 for the Lorenz system. For the full system of equations the correct model is rarely identified, only at the highest sampling rates with additional spurious terms. Typically SINDy will correctly identify the active terms for b, c, d and e equations but the a equation will contain many spurious non-zero terms. This is because the coefficients for the $\dot{a}$ equation are much larger than the other equations and so using one threshold for all equations results in the inclusion of terms that are not in the true model. Better results are seen by increasing the threshold for this equation. For the Lorenz system with a second order polynomial library the performance is better, returning models with lower coefficient error closer to the correct sparsities.

However, if we now compare the Lorenz system with a third-order polynomial library in Figure 61 (which contains 20 unknowns for each equation) to Knobloch system in Figure 59 (which constains 21 unknowns for each equation) we can see that the performance with 5 equations is better than a 3rd order system with a larger library. For the Lorenz system with a third-order polynomial library, the correct coefficients are effectively never identified. It appears in this case that either nonlinear correlations in the cubic terms or increased noise contributions due to cubic terms creates a harder regression problem than an increased problem dimension.

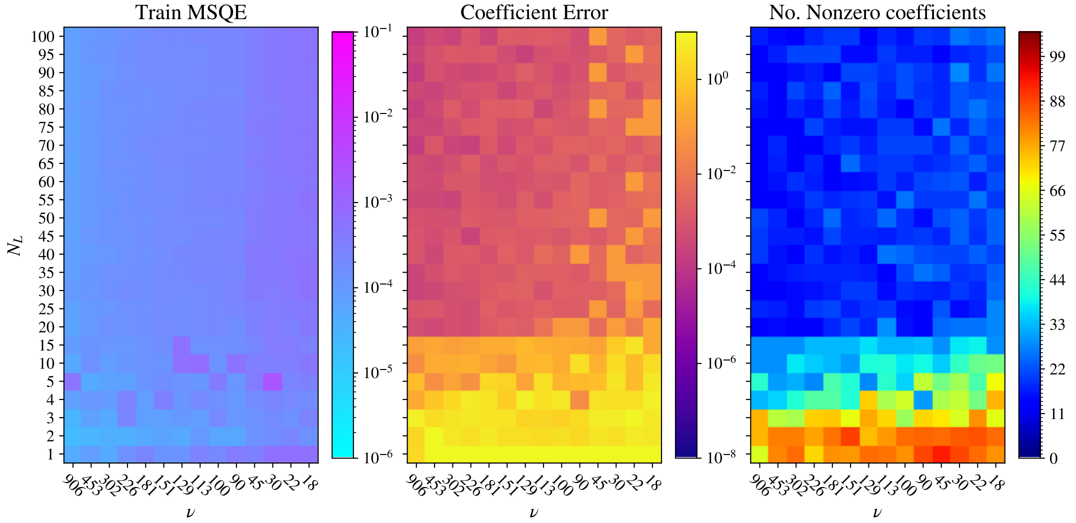


Figure 58: Variation of the total number of periods N_L of training data supplied against the sampling rate ν for a complete set of input time series a, b, c, d, e and noise with standard deviation 1%.

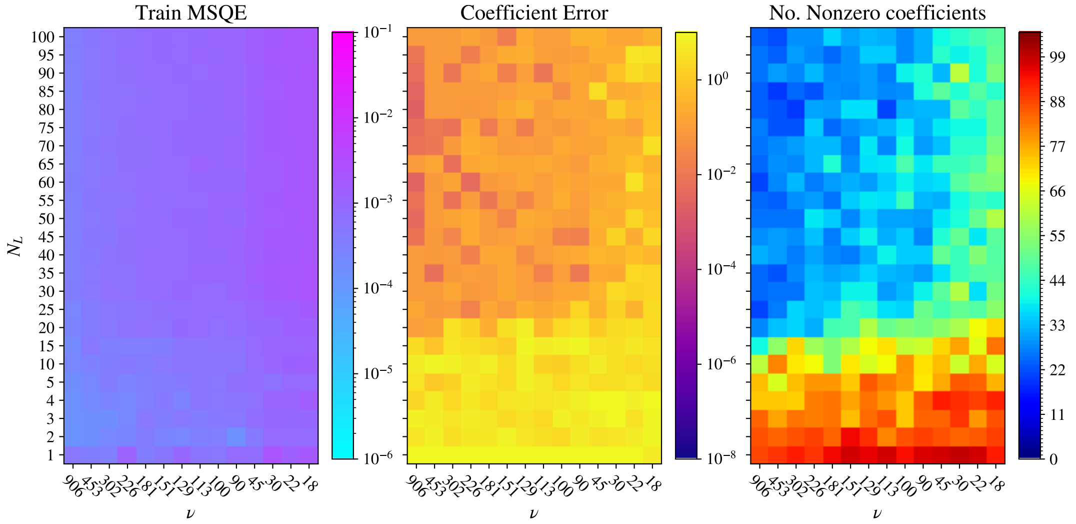


Figure 59: Variation of the total number of periods N_L of training data supplied against the sampling rate ν for a complete set of input time series a, b, c, d, e with a larger degree of noise having standard deviation 10%.

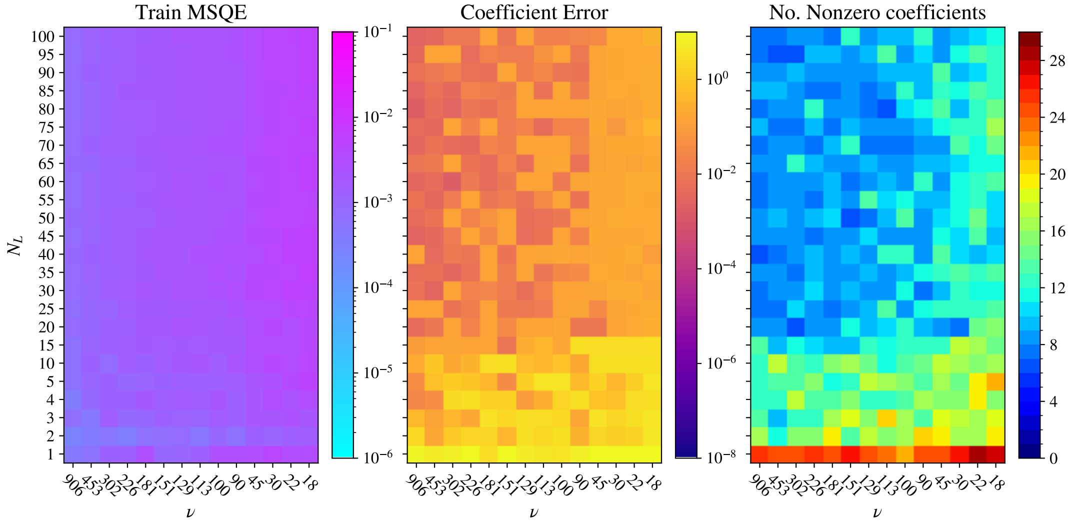


Figure 60: Variation of the total number of periods N_L of training data supplied against the sampling rate ν for a set of input time series restricted to only the Lorenz system a, b, c and noise with standard deviation 10%.

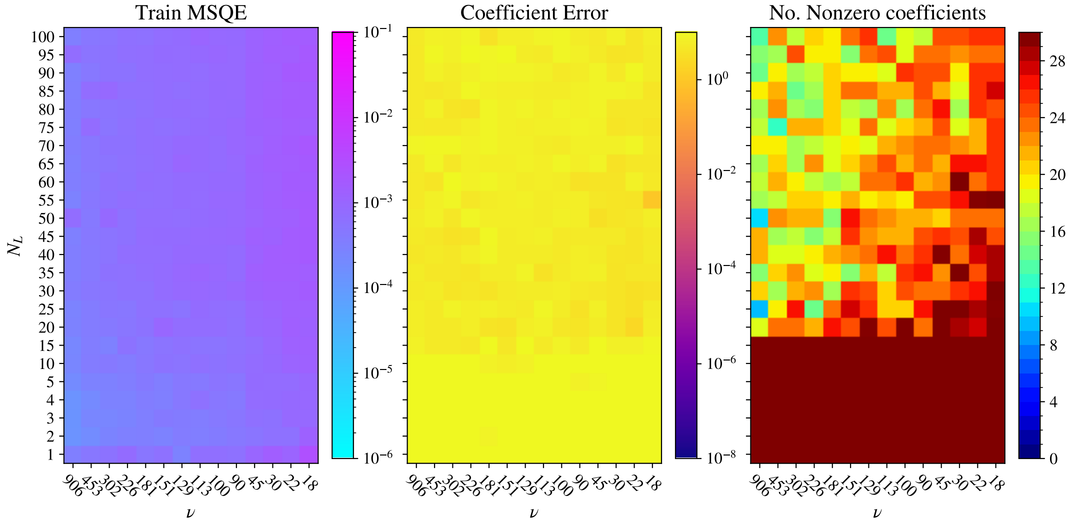


Figure 61: Variation of the total number of periods N_L of training data supplied against the sampling rate ν for a set of input time series restricted to only the Lorenz system a, b, c and noise with standard deviation 10% but with a larger third order polynomial library.

4.2.12 Modelling with prior physical knowledge

The Knobloch system has two properties which can be used generally in model identification [7]. The first is the symmetry

$$(a, b, c, d, e) \to (-a, -b, c, -d, e) \tag{4.68}$$

leaves the solutions unchanged. The second property relates to the divergence in the phase space

$$\frac{\partial \dot{a}}{\partial a} + \frac{\partial \dot{b}}{\partial b} + \frac{\partial \dot{c}}{\partial c} + \frac{\partial \dot{d}}{\partial d} + \frac{\partial \dot{e}}{\partial e} = -[Pr + (1 + \varpi + \zeta(5 - \varpi)] < 0 \tag{4.69}$$

so that solutions are attracted to a fixed point, limit cycle or strange attractor. The solution behaviour in Equation 4.68 places a set of equality constraints on the coefficients of the equations (for example terms linear in c cannot appear in $\dot{a}$). The constraints given by equation 4.69 place a series of inequality constraints on the sum of several coefficients at once, which can be deduced by simply constructing a polynomial library and then taking derivatives. For a second-order polynomial library, we can write a general library

$$\delta^i_0 + \delta^i_1 a + \delta^i_2 b + \delta^i_3 c + \delta^i_4 d + \delta^i_5 e + \delta^i_6 a^2 + \delta^i_7 b^2 + \delta^i_8 c^2 + \delta^i_9 d^2 + \delta^i_{10} e^2 + \\ \delta^i_{11} ab + \delta^i_{12} ac + \delta^i_{13} ad + \delta^i_{14} ae + \delta^i_{15} bc + \delta^i_{16} bd + \delta^i_{17} be + \delta^i_{18} cd + \delta^i_{19} ce + \delta^i_{20} de,$$

where the library is labelled for each equation i. In total we have a library of terms of this form for each equation. The phase space restriction places the following constraints on the library coefficients

$$\delta^a_1 + \delta^b_2 + \delta^c_3 + \delta^d_4 + \delta^e_5 < 0, \tag{4.70}$$

$$2\delta^a_6 + \delta^b_{11} + \delta^c_{12} + \delta^d_{13} + \delta^e_{14} = 0, \tag{4.71}$$

$$\delta^a_{11} + 2\delta^b_7 + \delta^c_{15} + \delta^d_{16} + \delta^e_{17} = 0, \tag{4.72}$$

$$\delta^a_{12} + \delta^b_{15} + 2\delta^c_8 + \delta^d_{18} + \delta^e_{19} = 0, \tag{4.73}$$

$$\delta^a_{13} + \delta^b_{16} + \delta^c_{18} + 2\delta^d_9 + \delta^e_{20} = 0, \tag{4.74}$$

$$\delta^a_{14} + \delta^b_{17} + \delta^c_{19} + \delta^d_{20} + 2\delta^e_{10} = 0. \tag{4.75}$$

which are found by taking the derivatives of the library and equating coefficients of common polynomial terms. The constraint outlined does not represent every possible model obeying this property and only restricts to a subset of possible models, but could still be beneficial to include. The symmetry constraints require

$$\delta^a_0 = \delta^a_3 = \delta^a_5 = \delta^a_6 = \delta^a_7 = \delta^a_8 = \delta^a_9 = \delta^a_{10} = \delta^a_{11} = \delta^a_{13} = \delta^a_{16} = \delta^a_{19} = 0, \tag{4.76}$$

$$\delta^b_0 = \delta^b_3 = \delta^b_5 = \delta^b_6 = \delta^b_7 = \delta^b_8 = \delta^b_9 = \delta^b_{10} = \delta^b_{11} = \delta^b_{13} = \delta^b_{16} = \delta^b_{19} = 0, \tag{4.77}$$

$$\delta^c_1 = \delta^c_2 = \delta^c_4 = \delta^c_{12} = \delta^c_{14} = \delta^c_{15} = \delta^c_{17} = \delta^c_{18} = \delta^c_{20} = 0, \tag{4.78}$$

$$\delta^d_0 = \delta^d_3 = \delta^d_5 = \delta^d_6 = \delta^d_7 = \delta^d_8 = \delta^d_9 = \delta^d_{10} = \delta^d_{11} = \delta^d_{13} = \delta^d_{16} = \delta^d_{19} = 0, \tag{4.79}$$

$$\delta^e_1 = \delta^e_2 = \delta^e_4 = \delta^e_{12} = \delta^e_{14} = \delta^e_{15} = \delta^e_{17} = \delta^e_{18} = \delta^e_{20} = 0, \tag{4.80}$$

Both of these constraints types can be used with the MIOSR optimiser, though the user must alter the pysindy code for MIOSR to allow inequality constraints. One advantage of using constraints with MIOSR is that it generally results in a performance increase, as fewer solutions need to be explored [160]. In the following we set $K = 1000$ to ensure the series is well sampled and set $H_{xt} = 0.906/3$. We set $\alpha = 1 \times 10^{-12}$ as generally it is observed that small α produces better results. Finally the target sparsity is set to 12, meaning that the total number of non-zero terms that will be identified over all equations cannot exceed this. This is the expected number of non-zero coefficients when $q = 0$.

The results of three cases are calculated. The first case in Figure 62 we consider the equivalent situation to STLSQ with 1% noise. With MIOSR, it appears that the coefficient error is actually higher than STLSQ. On the whole, most terms are correctly identified but solving the system in block diagonal form makes the system more sensistive to scaling issues of the time series. As the series for a is larger in amplitude, the derivatives corresponding to $\dot{a}$ are also larger. Optimisation methods which solve MSQE residuals are known to be sensitive to scaling issues, and the solution places emphasis instead on fitting the series with the largest contributions to the objective function. STLSQ does not suffer this issue as it solves each system iteratively. One solution to this is to scale the input time series but we then expect a corresponding scaling of the model coefficients.

Another alternative is to scale the feature library itself, as is done by [104] but this invalidates the developed inequality constraints. Figure 63 shows the same results for 1% additive noise when the input time series are scaled by their standard deviation. In this case the coefficient errors are much lower highlighting general scaling issues that can be encountered.

Figure 64 shows the case with 10% noise and no constraints, and Figure 65 shows the same case with constraints. We can see that there is comparatively little difference between the constrained and unconstrained cases in terms of coefficient error with noise. We also see that MIOSR is relatively insensitive to changes in sampling rates and total data length. Despite the relatively high coefficient error, we will discuss a practical example of using constraints in the following section.

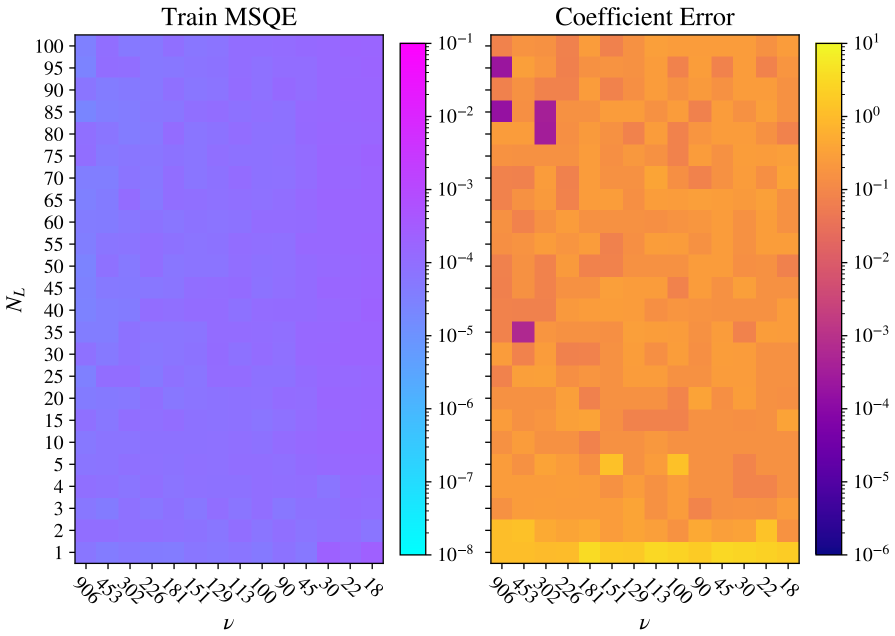


Figure 62: Train error and coefficient error while carrying the total length of training data N_L and the sampling rate ν with the MIOSR optimiser with 1% noise added to the state variables.

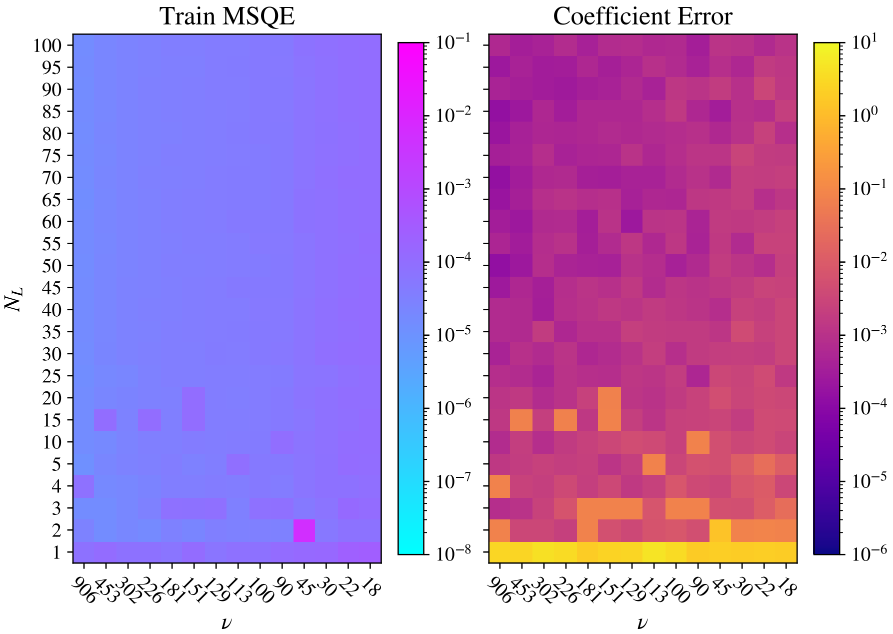


Figure 63: Train error and coefficient error while carrying the total length of training data N_L and the sampling rate ν with the MIOSR optimiser with 1%. In this Figure the input time series have standard deviations normalised to 1.

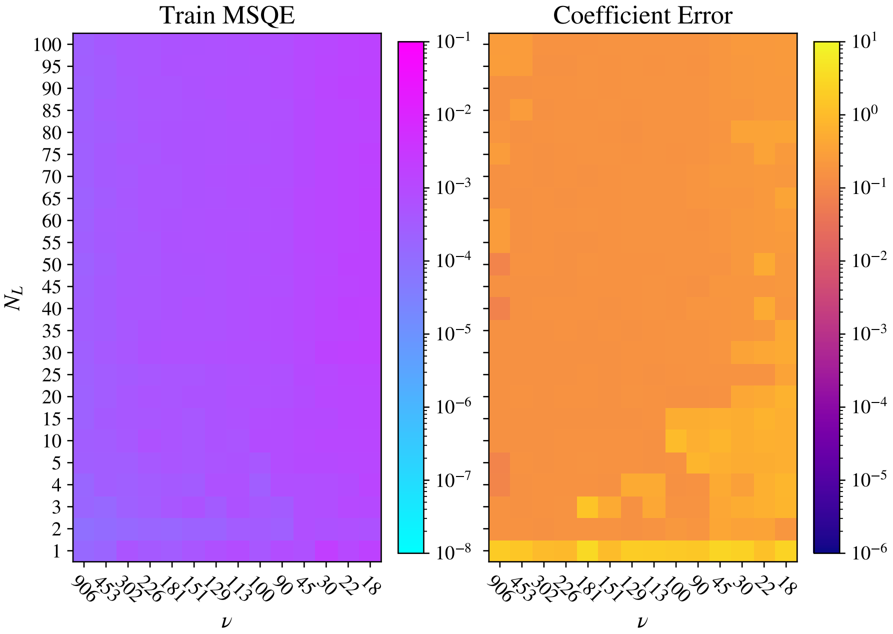


Figure 64: Train error and coefficient error while carrying the total length of training data N_L and the sampling rate ν with the MIOSR optimiser with a larger degree of noise at 10%.

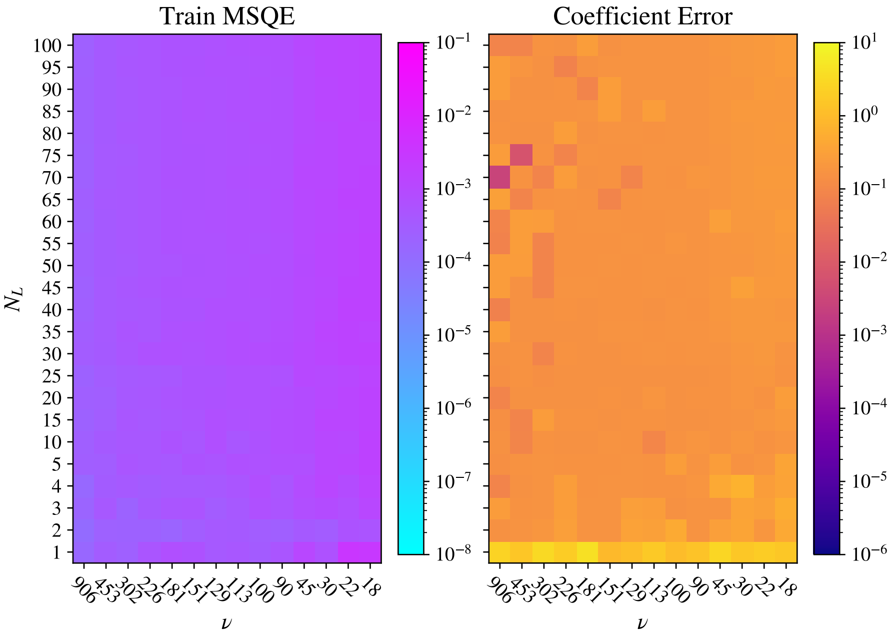


Figure 65: Train error and coefficient error while carrying the total length of training data N_L and the sampling rate ν with the MIOSR optimiser with 10% additional and also using constraints.

4.2.13 Explicit example

While the coefficient error does not particularly improve with the inclusion of constraints, we can consider an example where the data is subject to 20% noise with a sampling rate of $\nu \approx 20$ samples per Lyapunov time for the full 5 modes of the decoupled Lorenz system. This is a challenging case as the input data is highly corrupt by both noise and appreciably low sampling rates. We perform identification over 50 Lyapunov times and set $K = 10000$ with a target sparsity of 12. In Figure 67 we compare the phase space of the identified models showing the training data in blue, the constrained model in red and the non-constrained model in blue. Despite producing similar coefficients errors, the model found in the constrained case successfully reproduces the on attractor dynamics whereas the unconstrained model transitions to steady convection. For a test trajectory shown in blue in Figure 67, we start with larger initial conditions. Again, while the constrained model incorrectly models the transients, it still identifies a model which eventually converges to the attractor dynamics. In the unconstrained case this does not happen. This highlights a general issue with identification metrics: coefficient error may suggest when the true model is recovered, but other arguably suitable results can be found with terms which do not appear in the underlying equations.

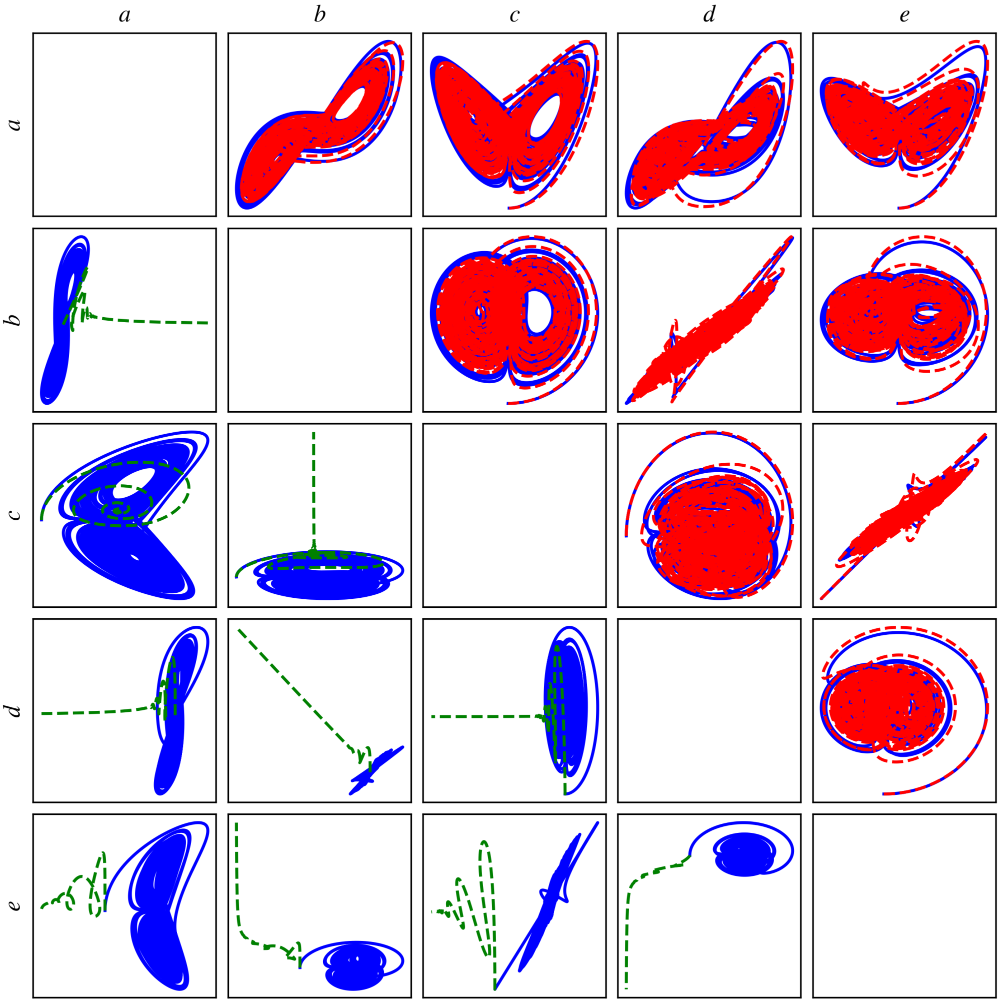


Figure 66: Phase space for the training data (blue), constrained model (red) and unconstrained model (green) found using the MIOSR optimiser.

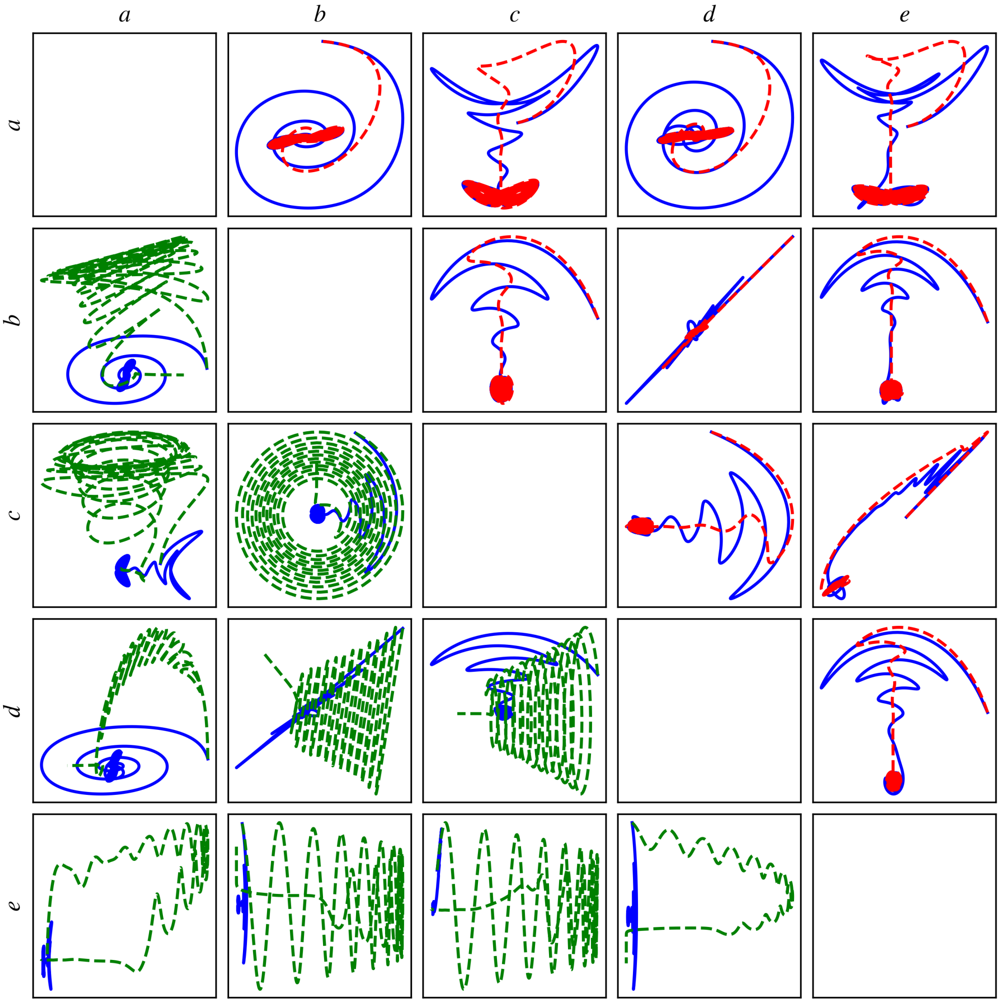


Figure 67: Phase space for the test data (blue), constrained model (red) and unconstrained model (green) found using the MIOSR optimiser.

4.3 Conclusions

We have considered some different cases of the performance of SINDy with different solution behaviours of the system given by Equation (4.45) - (4.49). The first case was overstable oscillations which proved challenging for SINDy. Even minor degrees of noise could prevent successful identification, the challenge resulting from correlation of the input library. Larger amplitude initial conditions helped this by separating the role of different nonlinearities, but even small degrees of noise caused large finite difference errors. If such cases are relevant for system identification, either constraining the library or the methods outlined by [154] could provide solutions to this. Generally though we conclude that SINDy is not suited to the overstable case when the dynamics are simplistic. Even when simpler models exist to explain the data (i.e., sparser models), SINDy still selects models with many correlated contributions. In this context the model sparsity could be limited either by increasing α or using the MIOSR optimiser. However, increasing α was already shown to produce identification of the incorrect model.

We then looked at the role of the nonlinearity of the oscillations themselves by performing identification over several different trajectories along an overstable branch before a transition to steady convection. The nonlinearity of the oscillations did not appear to change the suitability of different learning parameter substantially, and the main differences that arose were again related to errors in the derivative estimates. For fixed learning parameters however, increasing Rayleigh number trajectories were more robustly identified at lower sampling rates. When considering period doubling, SINDy also showed no improvement and almost entirely failed in the noiseless case where sampling rates were still arguably high. Again this is related to the reliance on derivative estimation from finite differencing. While constraints could be included in these cases, we only considered a second order polynomial library containing 20 unknown terms for each equation. The expectation of SINDy is to perform feature selection from a large available model space, and in this case fails even with noiseless data and a comparatively small library.

While robustness issues due to noise and sampling rates can be addressed with filtering and upscaling the data, the weak formulation of SINDy addresses both of these issues and is seen to perform better than the traditional formulation even on noiseless data [163]. When using the weak form we must make a selection of the integration domain size, but this can be related to physical time-scales of the system under consideration. In general it seems natural to employ a method which avoids taking derivatives entirely when sampling rates and noise are of concern. Some care must be taken though as the weak form can act as a high-frequency filter [152] and suitable selection of natural length-scales must be assessed beforehand. The weak formulation is shown to be substantially more robust and is effectively recommended to be used in all applications [163].

Applying the weak form, we showed that the performance of weak SINDy was actually better the conventional SINDy, even when high quality derivatives were provided to SINDy. The weak form was capable of reaching comparative sampling rates are SINDy, and also identified the underlying models with less data. In the case of the decoupled Lorenz system, the weak form was also quite robust to noise with only the Lorenz modes. Implementation of MIOSR in block diagonal form had some limitations due to scaling issues, but developed constraints improved the physicality of resulting models in poorly sampled noisy cases, even though the true model was not identified.

The assessment of the different cases considering provides insight into the expected performance of SINDy with experimental diagnostics on MAST-U. Many of these diagnostics show high degrees of noise [166, 174] and are typically quoted as having accuracies of ±10% of the signal standard deviation [111]. Further, while sampling rates can afford approximately 60 samples per period in the magnetics diagnostics, in other cases sampling rates are appreciably lower than this [26]. For any application to experimental time-series it therefore seems essential to use the weak form of SINDy, with the use of constraints also being desirable. One conceptual issue with experimental signals is not knowing if there are full or partial measurements. We considered one instance here where SINDy is capable of reproducing on attractor dynamics from incomplete measurements, however it is certainly not generalisable as it identifies the incorrect modal dependencies.

While weak SINDy provides a resolution to both noise and sampling rates, some consideration must be given to the possible difficulties faced by multiscale systems such as the ANAET model. As the ANAET model is a surrogate for expected tokamak behaviour, we can therefore anticipate issues arising when attempting to fit multiscale data. For conventional SINDy, successful identification will clearly be limited by the fastest time-scales present in the system. As mentioned, we have already identified multiple instances in which marginally poorer sampling rates prevent successful system identification with conventional SINDy. While we may hope to resolve sampling issues and derivative estimations from conventional SINDy by application of the weak form, we have seen that selection of an appropriate integration window is key. For multiscale system we will still be limited by the shortest time-scales in the system and therefore not take full advantage of noise smoothing from larger windows. We have also seen at least one instance where including a higher-order polynomial library caused poor system identification with the Lorenz equations compared to a comparable second-order library formed from all 5 equations. Given that the ANAET model captures diffusive effects with a term of the form $\dot{a}a^6$, we may expect to need a 7^{th} order polynomial library. This will create a challenging regression problem for SINDy. However, consideration of the imposed symmetries in the ANAET model would at least reduce the library complexity.

5 SINDy with magnetoconvection PDE data

In this chapter we use SINDy to derive sparse models from PDE simulations of the magnetoconvection problem described in § 4.1. In § 4.1, we derived a nonlinear coupled 5^{th} order system of ODEs given by equations (4.45)-(4.49). One limitation of these equations is that while they potentially offer a qualitatively valid description of the full PDE system, quantitatively they may only be valid close to the onset of instability under the assumptions with which they are derived. This situation presents an interesting testing scenario for SINDy as close to the onset of instability we know the system of ODEs will be derivable using SINDy. Further from the bifurcation point the weakly nonlinear theory will likely no longer quantitatively represent the dynamics, but we might then use SINDy to find alternative sparse models to describe the data.

To construct ODEs from PDE data, the first issue that must be addressed with SINDy is dimensionality reduction. Given that the size of a polynomial feature library grows combinatorically with the number of input time series, reducing the PDE data to a handful of time series is particularly important. For this we make use of POD, discussed in § 3.5.2. The reason this is useful is that POD can be thought of as a separable basis decomposition, where a function of the form is assumed separable and written in the form $f(x, t) = \phi(t)\psi(x)$. This decomposition is the same general type as the Fourier decomposition which is used to derive the low-order model given by equations (4.45)-(4.49). However, although close to the onset of instability we should expect the POD spatial modes to resemble Fourier modes, the spatial modes of the POD are constructed to optimally represent the data as described in [35], so will likely differ further from the bifurcation point. Moreover, there is no longer any reason why we should still derive something approaching the simplicity of the truncated nonlinear system. This makes for a relevant test case, as in many real applications we would have no guarantee that the input time series admit a sparse model, or that the choice of library is complete, or that the POD model suitably reproduces key features, such as for example, boundedness of solution.

5.1 Dedalus Simulations

Simulations of the magnetoconvection PDE system given by equations (4.24)-(4.26) with the stress-free boundary conditions are carried out using Dedalus [130]. Dedalus is an open-source spectral solver which allows symbolic equation entry. As we have written the system of equations in perturbative form, the imposed boundary conditions allow us to make use of a parity basis available in Dedalus (this was available in Dedalus 2 at time of writing, but not Dedalus 3). For instance, the boundary condition

$$T = 0, \quad z = 0, 1 \tag{5.1}$$

implies that the solution basis for T must be a sine basis in the z direction as the perturbations are zero at the upper and lower boundaries. Similarly the condition

$$\frac{\partial T}{\partial x} = 0, \quad x = 0, \lambda \tag{5.2}$$

implies that we must have a cosine basis as there is constant heat flux across the boundary (where sine has a non-zero derivative at the boundary). By writing the equations in perturbative form using a parity basis, we avoid the need to implement a Chebyshev basis in the z direction for the non-zero boundary conditions which typically results in slower run-times. This also means all results are computed on a uniform grid. For temporal discretisation we implement a 2nd-order semi-implicit BDF scheme included in Dedalus [66]. At each time-step Dedalus allows the user to save different outputs from the simulation. In our case, in addition to saving the state variables we also compute the velocities from the stream-function and the magnetic field from the flux function. Dedalus computes these in Fourier-space and therefore reduces the error in their computation. Each simulation output is saved at a time-step of $dt = 0.001$ to a final time of $T = 70$. All simulations on the oscillatory branch are carried out using an equally spaced grid of 96×96 points in the x and z directions. For the chaotic solutions, a resolution of 128×128 is used. The simulations for this Section were performed on the University of Leeds ARC4 facility, hosted and enabled through the ARC HPC resources and support team at the University of Leeds.

5.2 SINDy applied to Overstable oscillations

We first discuss applying SINDy to POD modes found from decompositions of magnetoconvection PDE data when the solutions display overstable oscillations. Each simulation is carried out with the parameters listed by ref [8], Figure 3b), with $Pr = 1$, $\zeta = 0.2$ and $Q = 500$ fixed for which linear theory predicts the onset of overstable oscillations at the Rayleigh numbers

$$R^{(o)} = 2306, \qquad r^{(o)} = 2.9598 \tag{5.3}$$

and the transition to steady convection at

$$R^{(e)} = 10649, \qquad r^{(e)} = 13.6651. \tag{5.4}$$

Ref. [8] finds that along this branch, the transition to steady convection occurs significantly below $r^{(e)}$.

5.2.1 Normalisations of the POD modes

When comparing the POD temporal modes to the weakly nonlinear model, care must be taken. A POD decomposition finds modes which satisfy

$$\boldsymbol{U}^T\boldsymbol{U} = \sum_i U_{ij}U_{ij} = \mathbb{I} \tag{5.5}$$

and the modes are therefore orthonormal on the discrete inner product. In cases where the weakly nonlinear model is valid, we expect the identified POD modes to resemble the weakly nonlinear modes. To compare the modes, we must first consider the normalisation differences between the two. The modal expansion in the weakly nonlinear model in ref. [7] comprises of orthogonal but not orthonormal modes. To make comparisons between the POD modes and weakly nonlinear modes, we must identify the appropriate scalings for the POD modes so that they are normalised correctly. The inner product is defined

$$\langle \phi_i, \phi_j \rangle = \int_0^1 \int_0^\lambda \phi_i \phi_j \mathrm{d}x\mathrm{d}z = \delta_{ij}. \tag{5.6}$$

The modes given by ref [7] are as follows

$$\psi(\boldsymbol{x}, t) = 2^{3/2} p \frac{\lambda}{\pi} a(\tau) \sin(\pi x/\lambda) \sin \pi z, \tag{5.7}$$

$$T(\boldsymbol{x}, t) = 1 - z + 2(2/p)^{1/2} b(\tau) \cos(\pi x/\lambda) \sin \pi z - \frac{1}{\pi} c(\tau) \sin 2\pi z, \tag{5.8}$$

$$A(\boldsymbol{x}, t) = x + 2(2/p)^{1/2} \lambda d(\tau) \sin\left(\frac{\pi x}{\lambda}\right) \cos \pi z + \frac{\lambda}{\pi} c(\tau) \sin\left(\frac{2\pi x}{\lambda}\right) \tag{5.9}$$

where $\tau = pt$ and $p = \pi^2(1 + 1/\lambda^2)$. The transformation between the weakly nonlinear modes and orthonormal modes (denoted by $'$) is given by

$$a' = \sqrt{\frac{\lambda}{4}} 2^{3/2} \lambda \frac{p^{1/2}}{\pi} a(\tau), \tag{5.10}$$

$$b' = \sqrt{\frac{\lambda}{4}} \left(\frac{2^{3/2}}{p^{1/2}} \right) b(\tau), \tag{5.11}$$

$$c' = \sqrt{\frac{\lambda}{2}} \frac{1}{\pi} c(\tau), \tag{5.12}$$

$$d' = \sqrt{\frac{\lambda}{4}} \left(\frac{2^{3/2}}{p^{1/2}} \right) d(\tau), \tag{5.13}$$

$$e' = \sqrt{\frac{\lambda}{2}} \frac{1}{\pi} e(\tau). \tag{5.14}$$

For all the results, we will work in the normalisations given by ref [7]. In other words, we identify modes associated to the primed variables and transform to the unprimed variables. The main advantage is that the series are normalised by the critical Rayleigh number in the field-free case and therefore many of the coefficients have similar orders of magnitudes. It also allows comparisons of the POD decomposition directly with the weakly nonlinear model. There is one further normalisation that must be considered coming from the difference between the continuous and discrete inner products. Each POD mode is also multiplied by

$$N = \sqrt{\frac{1}{N_x N_z}}, \tag{5.15}$$

where N_x and N_z are the number of points in the x and z directions respectively.

5.2.2 Equation recovery metrics

We now discuss some chosen model selection metrics, noting complete discussions have already been given in §3.2. We use two different metrics which are used for benchmarking in ref. [163], with extensive explanations given by [105] in relation to applications with SINDy. The first metric is simply the mean-square error of the predicted derivatives, given by

$$MSQE = \frac{1}{nm} \sum_{i=1}^{m} \sum_{j=1}^{n} (\dot{X}_{ij} - \dot{Y}_{ij})^2 \tag{5.16}$$

where $\dot{\boldsymbol{X}}$ is the matrix of true derivatives and $\dot{\boldsymbol{Y}}$ is the matrix of predicted derivatives from the identified SINDy model. The MSQE simply gives a measure of how well the model fits the objective, but in many cases gives no estimate of overfitting. Cases with low error can be obtained when models are overfitted, particularly when the series is noisy. The second model selection metric is the AIC score used by ref. [105]

$$AIC = m \log(||\dot{\boldsymbol{X}} - \dot{\boldsymbol{Y}}||_2^2) + 2k + \frac{2k}{m - k - 1} \tag{5.17}$$

where $|| \cdot ||_2^2$ is the $L2$ norm and k is the number of non-zero coefficients over all equations. The optimal model is then the model which minimises the AIC score. The AIC score favours sparser models by penalising models with many non-zero coefficients (larger k). As the AIC score can be arbitrarily negative, often the relative AIC score is calculated instead

$$AIC_c = AIC - AIC_{min}. \tag{5.18}$$

With the relative AIC score, only one model will have the value 0. We use the definition given by [163] but note that the chosen likelihood function minimises the $L2$ norm which is not normalised by the total number of points. This does not matter in cases where the total number of points is the same. For each POD data-set, we use the first 80% as a training set, and the final 20% as a test set.

5.2.3 General POD results

POD is performed over data-sets along the oscillatory branch with fixed $Q = 500$, $\zeta = 0.2$, $Pr = 1$ and Rayleigh number varying $R \in [2400, 2500, 3000, 4000, 5000, 6000, 7000]$. POD is performed on each field and each data-set separately, from which the spatial modes and corresponding time-series are found. The results for the POD decomposition are shown in Figure 68 along with the corresponding time series for the evolution of ψ_0 which is labelled a in blue, with a comparison of the weakly nonlinear model integrated at the same parameter values initialised from the first point in the time series corresponding to the POD decomposition. When $R < 4000$, the POD decomposition closely agrees with the weakly nonlinear model, finding spatial modes which resemble Fourier modes. Further from the bifurcation point, the weakly nonlinear model transitions to steady convection earlier than predicted by the simulations. After the transition to steady convection the identified POD modes no longer resemble Fourier modes, so no comparison can be made between the amplitude of the convection to the weakly nonlinear model.

The corresponding singular values are shown in Figure 69 for each of the simulated Rayleigh numbers. For $R = 2400$, most of the energy is captured in the leading mode for ψ, two modes for T and two modes for A. As R increases, the floor of the singular value spectrum also increases and the energy in the leading singular value also decreases until the transition to steady convection. For T and A, close to the transition of steady convection at $R = 6000$, the two leading singular values capture similar energy contents. In Figure 68, while the mode structure is qualitatively similar at each Rayleigh number, for A the leading modes swap after $R = 5000$. For training data related to $R = 5000$ and $R = 6000$, these two modes corresponding to A_0 and A_1 are swapped so that the modal decomposition resembles that of the weakly nonlinear model. This needs to be considered when enforcing symmetry constraints later.

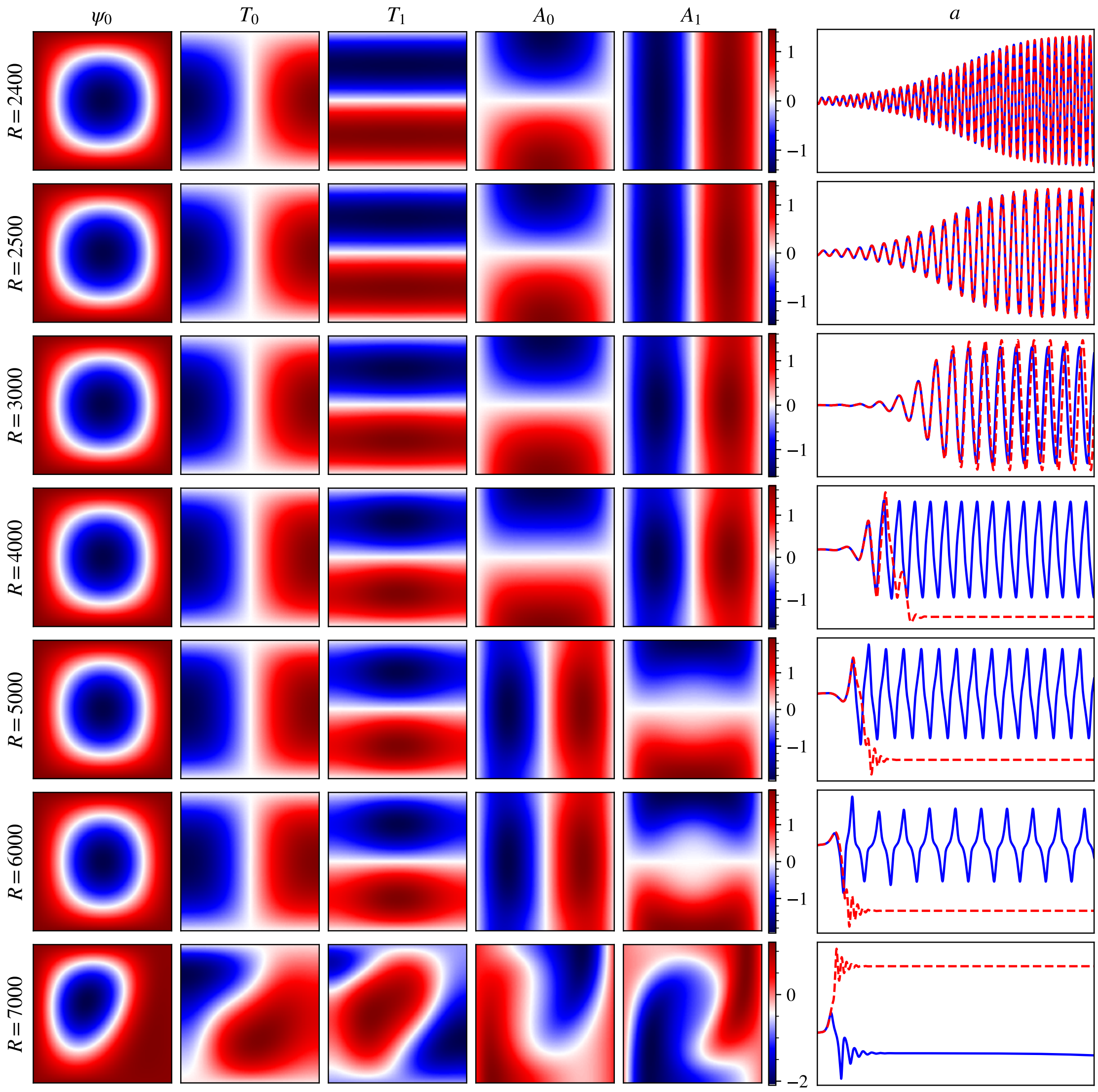


Figure 68: Comparison of the modes identified at each Rayleigh number on the oscillatory branch which transitions to steady convection after $R = 6000$. The right column shows the time series for ψ_0 in blue compared to the weakly nonlinear model in dashed red. Note that the end time is not the same and is set for clarity.

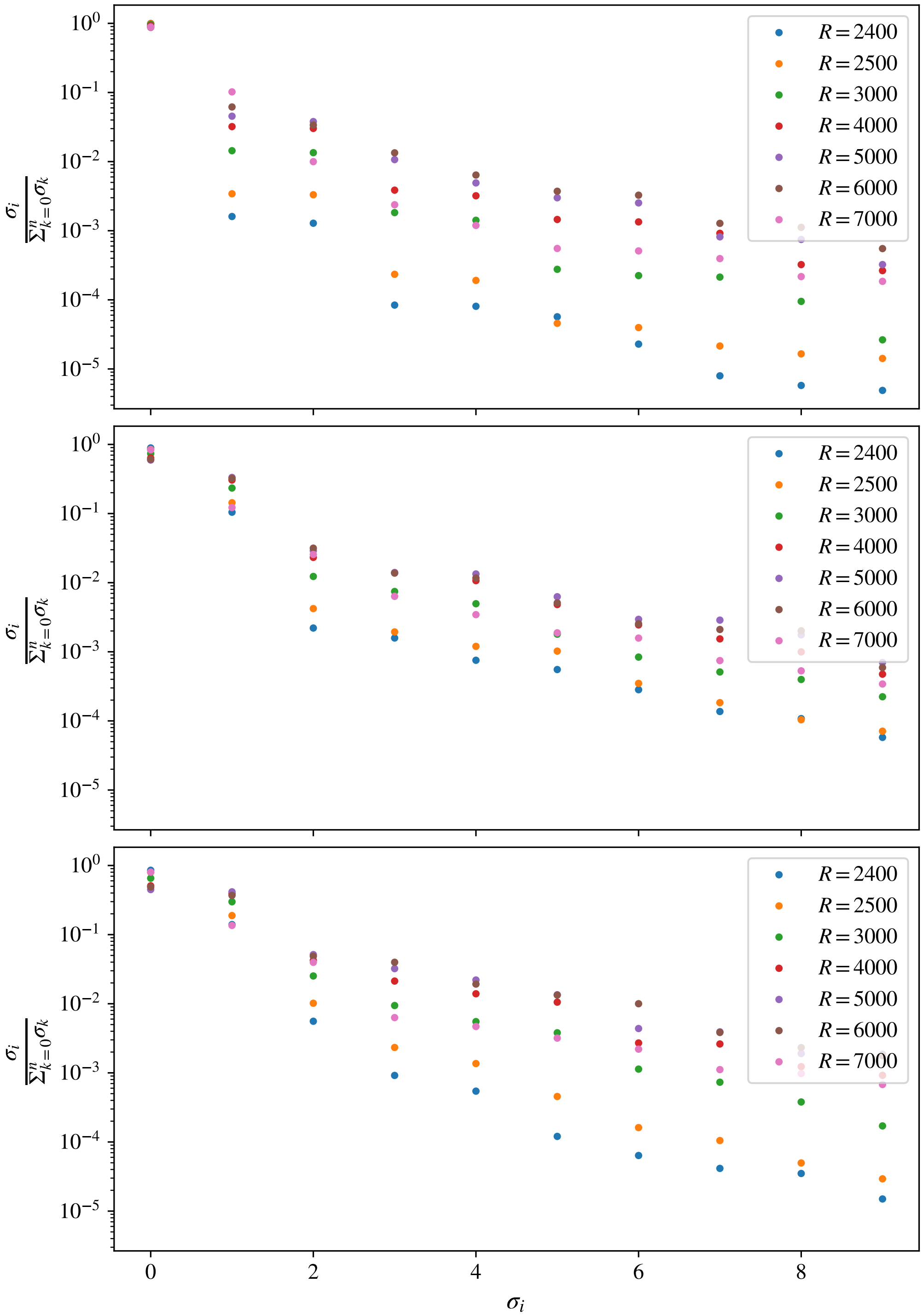


Figure 69: Comparison of the singular value spectrum for ψ (top), T (middle), A (bottom). The y-scale shows the relative energy captured in each singular value.

5.2.4 POD decomposition at the onset of oscillatory convection

Before discussing SINDy fits to different Rayleigh numbers on the whole, we first consider the case close to the onset of overstable oscillations where the POD decomposition agrees closely with integrations from the weakly nonlinear model. We take results from $R = 2400$ which can be seen to closely agree with the weakly nonlinear model in Figure 68. The singular value spectrum in Figure 69 also shows that the majority of the energy is captured within the first mode for ψ, the first two modes for T and the first two modes for A. For $R = 2400$, we also plot modes from the next singular-values in the spectrum for ψ in Figure 70. It can be seen that modes for other singular-values still resemble a Fourier basis. Similar results are seen for T and A.

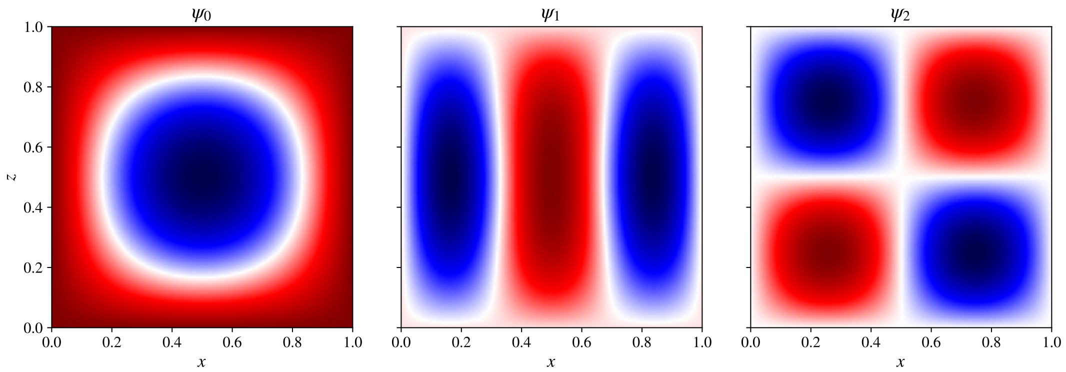


Figure 70: First three spatial modes for ψ identified from the POD decomposition.

5.2.5 System identification at $R = 2400$

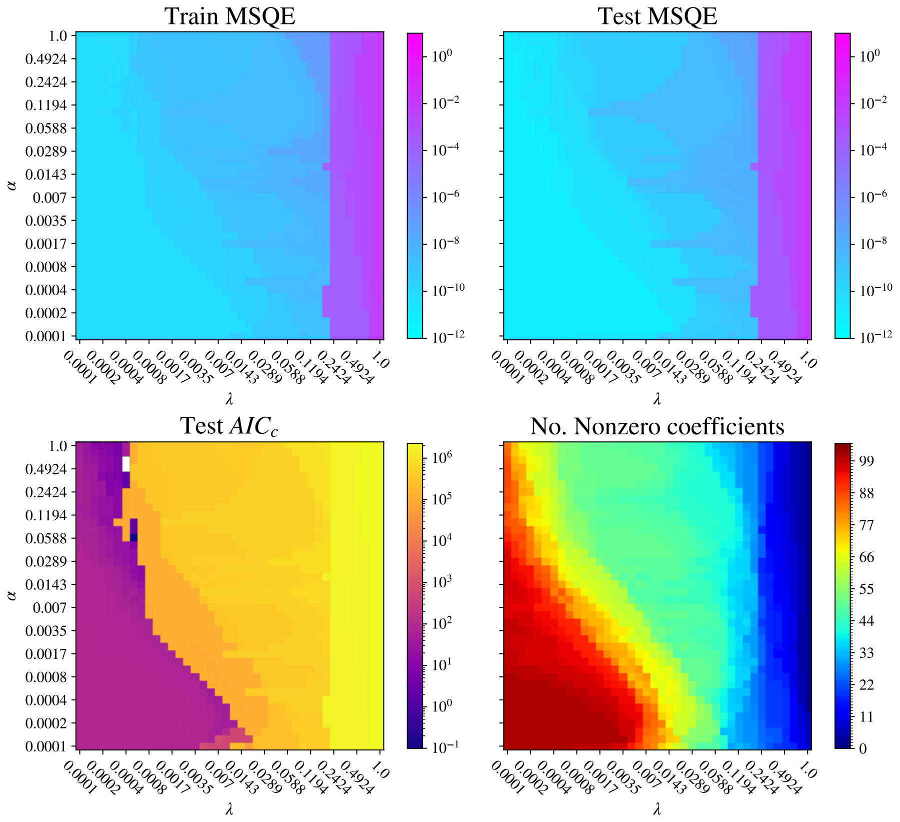


Figure 71: Calculated error while varying the L2 regularisation α and the coefficient threshold λ for $R = 2400$ close to the onset of overstable oscillations.

We first compute the aforementioned error metrics for $R = 2400$ while varying α and λ with results shown in Figure 71. We can see that both the test and train MSQE obtain very small values in regions of small α and small λ. Use of small α and λ results in cases which are inevitably overfitted as can be seen from the number of non-zero coefficients in these regions. The test MSQE does not provide any new information in this case as the test set is reserved from the end of the POD data when the oscillations are on attractor and hence the data is not novel. The entire region where the models have the most populated number of non-zero terms yields the same lowest MSQE.

The minimum test AIC_c score instead favours a model with small λ and larger α, indicated by white squares. This in itself is interesting, as the α corresponds to regularisation which improves regression with correlated features. The models here also have fewer non-zero coefficients, though they are still significantly less sparse than the intended weakly nonlinear model. In both cases if data is generated from the weakly nonlinear model at the given simulation time-step and identified learning parameters here, we can successfully recover the equations with the same active terms as the weakly nonlinear model. Further, if we constrain the feature library so that we perform a regression only on the active terms of the weakly nonlinear model, we recover the equations with the expected coefficient values. However, as mentioned in the previous section it is extremely sensitive to noise. Close to onset of instability, the modes are correlated and SINDy cannot easily distinguish between the correct features despite only using a second order polynomial library.

Integration of the resulting models produce good predictions for the on-attractor dynamics given by the POD modes. Again this is because while the model is not sparse, many terms effectively cancel out on the attractor (see appendix of ref [163]). The fact that terms effectively cancel out prevents STLSQ from finding the sparsest solution in this case, as the method thresholds by the size of coefficient value.

5.2.6 $R = 3000$ with larger initial conditions

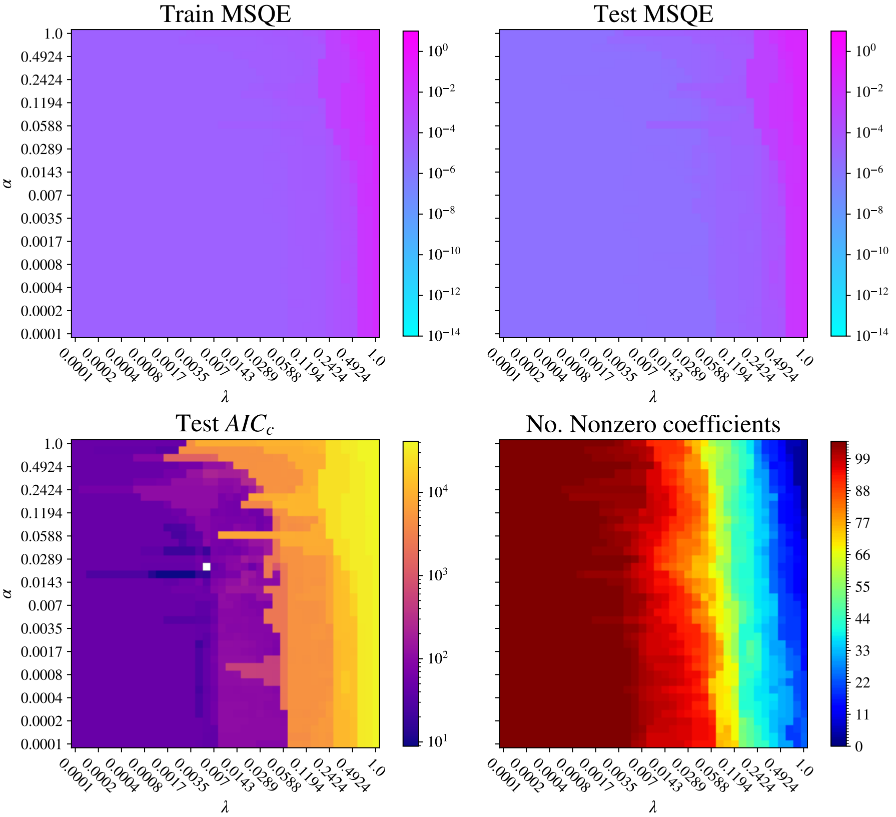


Figure 72: Calculated error while varying the L2 regularisation α and the coefficient threshold λ for $R = 3000$ when simulations are initialised from an $R = 8000$ steady convection simulation.

In the previous section it was noted that for the weakly nonlinear system, starting at larger initial conditions off the final attractor gave more information on the correct active nonlinearities. Including large amplitude transients made model identification more robust, so to assess this we initialise a simulation from the final time-step of an $R = 8000$ simulation during steady convection and continue the simulation at $R = 3000$. By doing so we no longer start from a small initial condition and hope to provide information on off-attractor dynamics which will make correct identification of the active nonlinearities simpler.

Figure 72 shows the selection of learning parameters for the larger initial condition. We see that despite starting from a larger initial condition there has been no substantial change in the sparsity of the model. The minimum AIC_c score again indicated by the white square still favours a model with a relatively low sparsity, with many more active terms than the weakly nonlinear model (14 terms). In this instance there is less dependence on α as the initial condition creates some decorrelation between modes and thus the problem benefits less from $L2$ regularisation. Despite this, it still appears that minor differences between the POD truncation and the Fourier basis in the weakly nonlinear model prevent successful identification. Indication of poor fitting is also given by the large training MSQEs which imply that the transient is not well captured by the library.

5.2.7 Model identification at $R = 6000$

While STLSQ has only been applied to a handful of cases, a principled approach to model selection is challenging to develop as in many cases non-sparse models are favoured. Further, the presence of correlation in the input time series at low Rayleigh number makes identifying sparse models challenging. The main solution is to make use of a large α but this inevitably favours the sparsity constraint over faithfulness to the data. For the remainder, we instead consider application of the MIOSR optimiser as we are able to control the allowable sparsity of the final solution. We are also able to implement developed constraints given in the previous section which may improve model identification by reducing the potential model space.

Before discussing parameter sweeps for all Rayleigh number cases, we first perform a constrained regression of the POD decomposition at $R = 6000$ before an observed bifurcation to steady convection at $R = 7000$ observed in the Dedalus simulations. The first regression we perform in this case, the feature library is restricted to only allow features in the weakly nonlinear model to appear. At this Rayleigh number, the weakly nonlinear model predicts steady convection contrary to the full numerical simulations. The weakly nonlinear model has the explicit form

$$\dot{a} = -a + 7.6995b - 2.533d(1 + (2.0 - 1)e)], \tag{5.19}$$

$$\dot{b} = -b + a(1 - c), \tag{5.20}$$

$$\dot{c} = -2c + 2ab, \tag{5.21}$$

$$\dot{d} = -0.2d + a(1 - e), \tag{5.22}$$

$$\dot{e} = -0.4e + 2ad. \tag{5.23}$$

A regression on to the POD modes with the exact form stated above gives the model

$$\dot{a} = -1.296a + 7.870b + 0.536d - 7.163de, \tag{5.24}$$

$$\dot{b} = 1.015a - 1.053b + 0.980ac, \tag{5.25}$$

$$\dot{c} = -2.015c - 1.905ab, \tag{5.26}$$

$$\dot{d} = -1.428a - 0.445d - 1.796ae, \tag{5.27}$$

$$\dot{e} = -0.429e + 0.815ad. \tag{5.28}$$

Many of the identified coefficients are very similar, except coefficients associated with the magnetic field (d and e). This system still predicts steady convection at the specified Rayleigh number, showing that the weakly nonlinear model will not be recovered at this Rayleigh number.

One possible improvement can be made by including high-order terms in the feature library. By expanding the library to contain cubic feature terms and increasing the total sparsity to 18 we identify a model of the form

$$\dot{a} = 4.098b - 1.059d - 2.091de - 7.767bce, \tag{5.29}$$
$$\dot{b} = 1.015a - 1.053b - 0.980ac, \tag{5.30}$$
$$\dot{c} = -1.693c + 1.938ab - 0.916bcd, \tag{5.31}$$
$$\dot{d} = -0.398e - 0.257ab + 1.234ad - 0.706abe, \tag{5.32}$$
$$\dot{e} = 1.525a - 0.851b - 1.408ae - 1.184bce. \tag{5.33}$$

This model has a similar form to the weakly nonlinear model, with 4 additional cubic terms added, selected from a total of 56 library terms for each equation. Comparison of the cubic model to the weakly nonlinear model at the given parameters is shown in Figure 73 which shows good agreement to the training trajectory when integrated from the same initial condition. If the coefficient for b is increased by $\tilde{r} \approx 1.28$ corresponding to an increase to $R = 7000$, the identified system predicts the onset of steady convection at $R = 7000$ which is seen numerically, shown in Figure 74. Note that no real comparison of magnitude can be made in this case as the POD decomposition is at a different Rayleigh number and therefore the identified basis is different for steady convection.

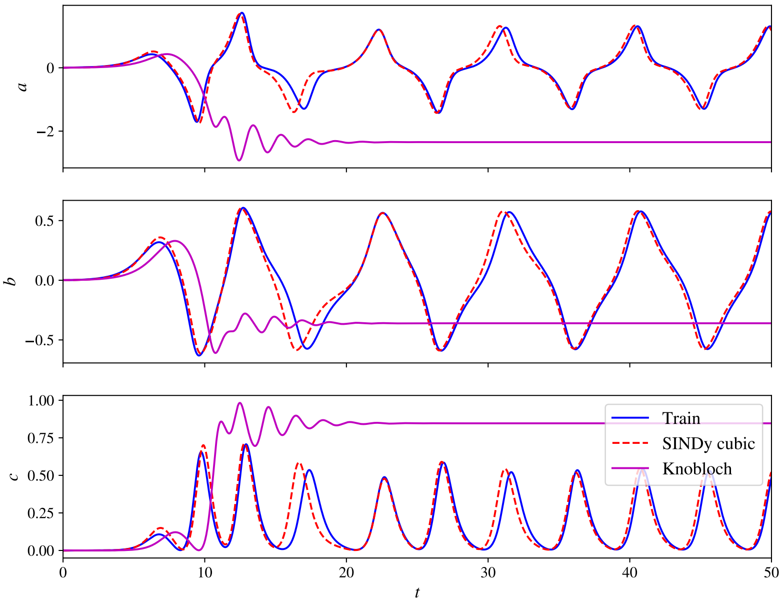


Figure 73: Comparison of the cubic model identified with the training trajectories and the weakly nonlinear model.

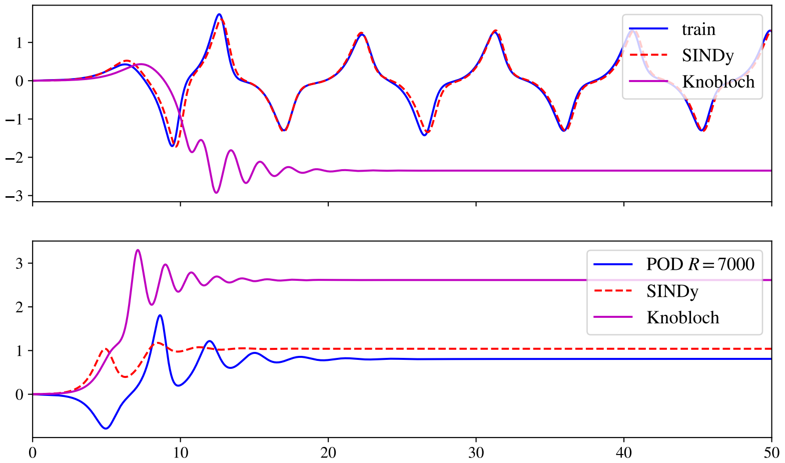


Figure 74: Comparison of the cubic library on the original training data (top) and the cubic model to the POD trajectories found at $R = 7000$ showing the model bifurcates to steady convection.

5.2.8 Pareto front for MIOSR derived models

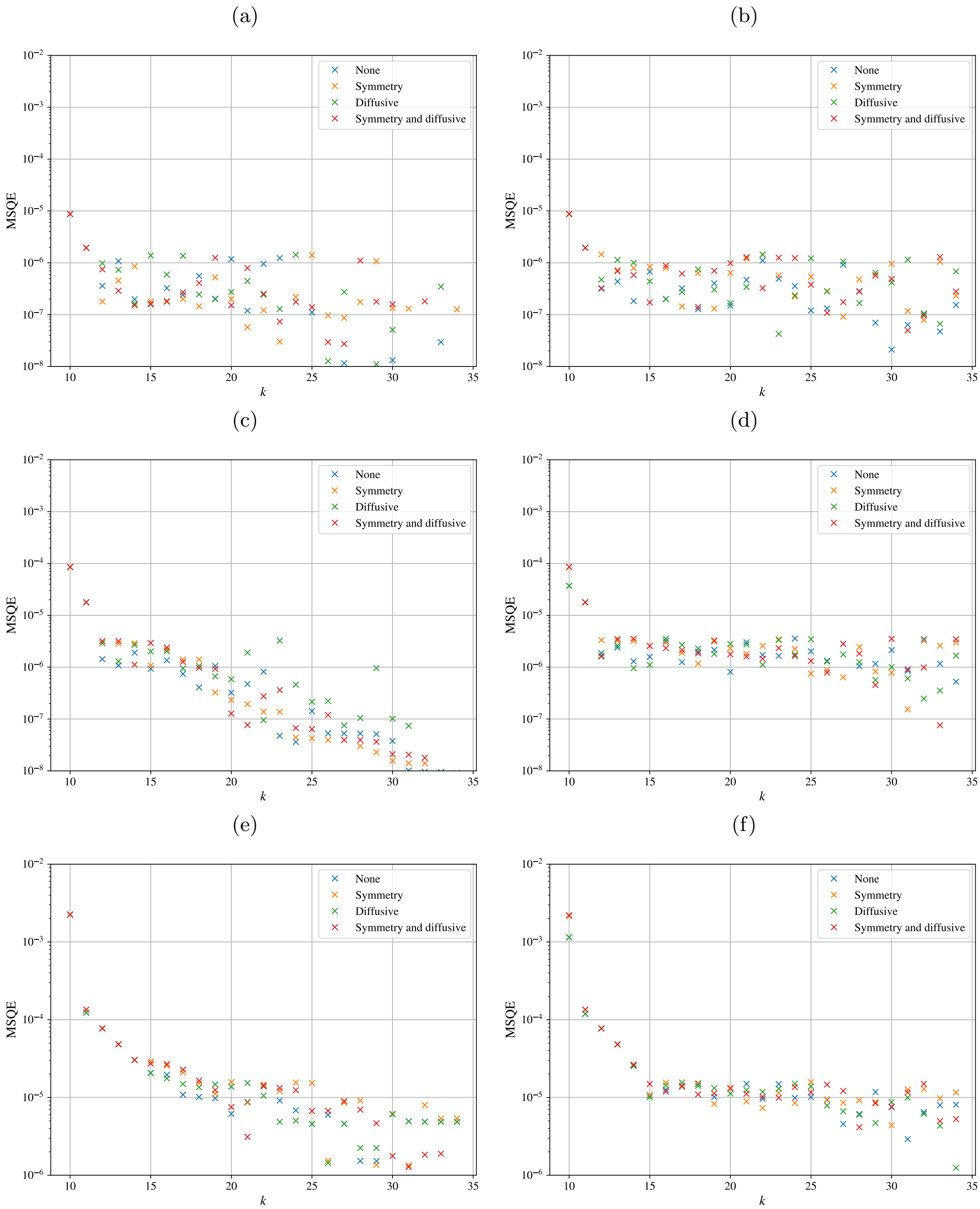
Inspection of a model found with MIOSR at $k = 18$ and $R = 6000$ suggests that we can use MIOSR to find sparser solutions than those given by STLSQ. Further, given that there is little dependence on α, we attempt to reduce the complexity of the problem by fixing $\alpha = 1 \times 10^{-10}$ and plotting the error curves for each Rayleigh number with both second and third-order polynomial libraries. For each of these cases we also consider fits using either no constraints, symmetry constraints, diffusive constraints or both symmetry and diffusive constraints, the details of which were discussed in §4.2.12. For the cases of $R = 5000$ and $R = 6000$ we recall that the modes for A are swapped as otherwise the symmetry constraints are not valid.

The results of the analysis are shown in Figure 75 for $R = 2400$, 2500, 3000 and Figure 76 for $R = 4000$, 5000, 6000. Both Figures show the error on the $y-$axis against an increasing number of coefficients k on the $x-$axis when either no constraints or some differing constraints are included. Considering first the results of Figure 75 we can see that close to onset of overstable oscillations at $R = 2400$ and $R = 2500$, both second and third-order polynomial libraries produce low scoring models (with MSQE below 10^{-5}) within model sparsities of 10-14. There is an elbow in the error around these model sparsities, though the error eventually continues to decrease with decreasing sparsity (increasing k). Again we identified this case in the ODE validation as finding models with many non-zero terms due to correlated input features. Decreasing sparsity will then continue to result in reduced error as the inclusion of more correlated features allows closer fit to the training data. Inclusion of both constraints or either symmetry and diffusive constraints has no appreciable improvement on the MSQE. However, given the correlated input features at $R = 2400$ we already know it is possible to find multiple models with low MSQE and so this is result is not surprising. As we increase Rayleigh number to $R = 3000$, the number of terms required to reach similar MSQEs increases, though slightly sparser models are required in the cubic case.

For higher Rayleigh numbers shown in Figure 76, we again see that higher sparsity models are favoured to produce MSQEs below 10^{-4}. We also see that third-order polynomial libraries produce lower scoring models compared to second-order polynomial libraries for the same sparsities. As we approach $R = 5000$ and $R = 6000$ the elbow in error becomes more pronounced as the temporal POD modes used for training SINDy become less correlated and more nonlinear. Results from the ODE evaluation suggest that identification should be more robust at higher Rayleigh number at the given sampling rates trained on here. For $R = 4000$ in the cubic case, diffusive constraints alone produce higher errors than all other cases. This is because diffusive constraints can make a more challenging optimisation problem and if the allowed regression time is increased then these errors again decrease to agree with the unconstrained case.

As both constrained and unconstrained models produce similar scores, we consider an exploration of the constrained models only in Figure 77 for the second-order polynomial library and Figure 78 for the third-order polynomial library. We compare integrations of identified model (in red) against a section of data reserved at the end of the POD time series (in blue). Models are compared for each Rayleigh number and varying model sparsities in the vertical. In producing the following Figure, we restrict the minimum allowable time-step to 10^{-5} for the integration solver to promote reasonable model integration times. If this restriction is not made model integration is too time consuming, especially when we are concerned with deriving simplified models of PDE behaviour. The restriction on the solver results in certain models growing to large values and this sets the extent of the plots (each subfigure does not share a common range in the x and y axes). Consequently, the training data in plots with poorly performing models can appear very small on the plot. We emphasise that the training data in every column is shared, regardless of appearance.

For the second-order polynomial library in Figure 77 we find promising fits for $R \leq 3000$ for a large choice of sparsities. However, for higher Rayleigh numbers the second-order polynomial library struggles to reproduce the dynamics for sparsities less than $k = 20$. While not shown here, fits with $k = 20$ do reproduce the dynamics for higher Rayleigh numbers. Conversely for the third-order polynomial library in Figure 78, the model fits are less consistent for decreasing sparsity, with some models of a given sparsity fitting well but an increase causing poorer fits. This is particularly true for lower Rayleigh numbers, with correlation in the input time series compounded by larger library sizes. However, for higher Rayleigh numbers the third-order library is capable of reproducing the dynamics at higher sparsities of $k = 17$ for $R = 6000$. The results of integration support the MSQE calculations which showed that denser models were required for the second-order library at higher Rayleigh numbers compared to those of the third-order polynomial library.


(a)

(b)

(c)

(d)

(e)

(f)

Figure 75: Calculated MSQE on the training derivatives for a),b) $R = 2400$, c), d) $R = 2500$, e), d) $R = 3000$. The left-hand column are results for a second-order polynomial library and the right-hand column are for a third-order polynomial library.

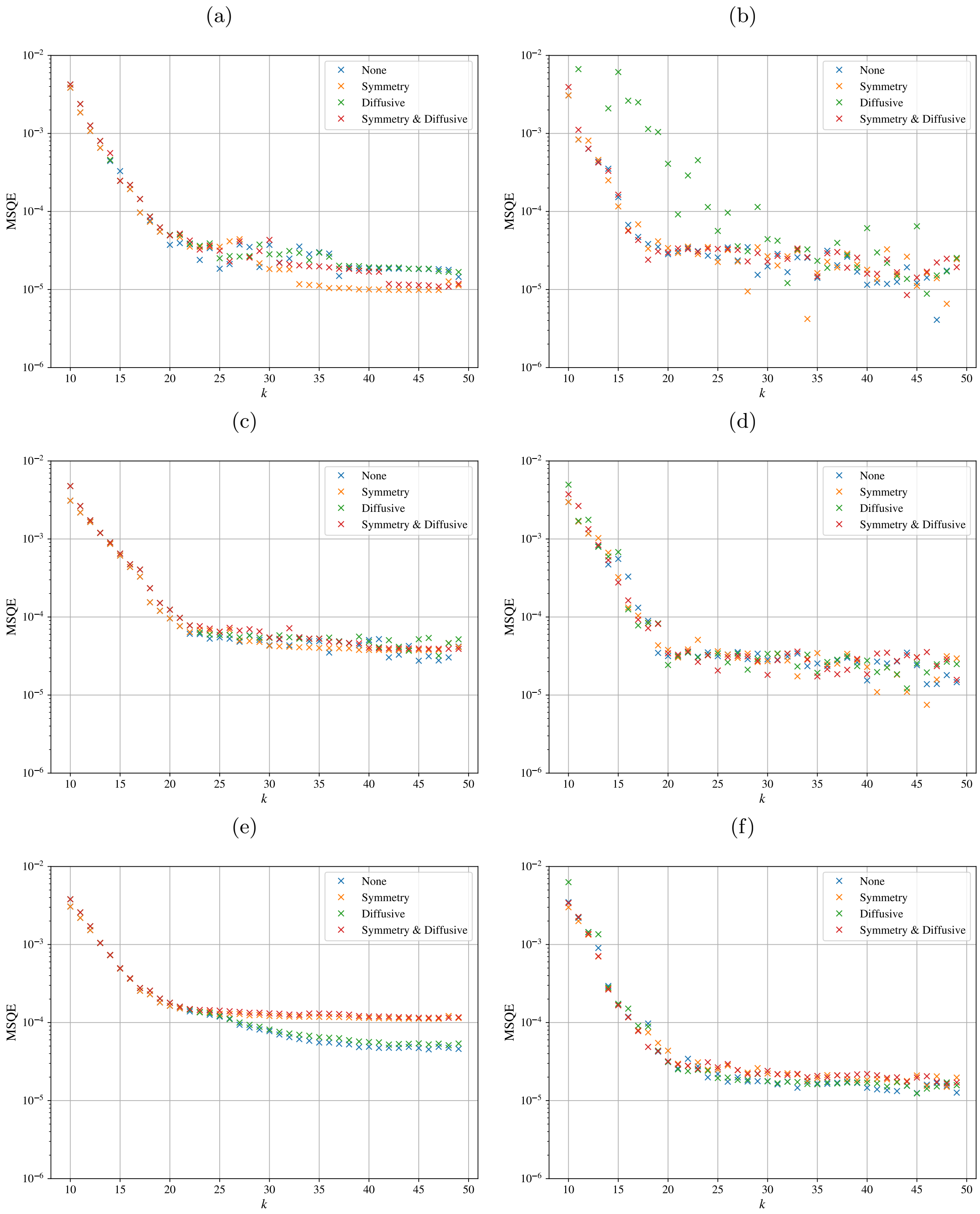


Figure 76: Calculated MSQE on the training derivatives for a),b) $R = 4000$, c), d) $R = 5000$, e), d) $R = 6000$. The left-hand column are results for a second-order polynomial library and the right-hand column are for a third-order polynomial library.

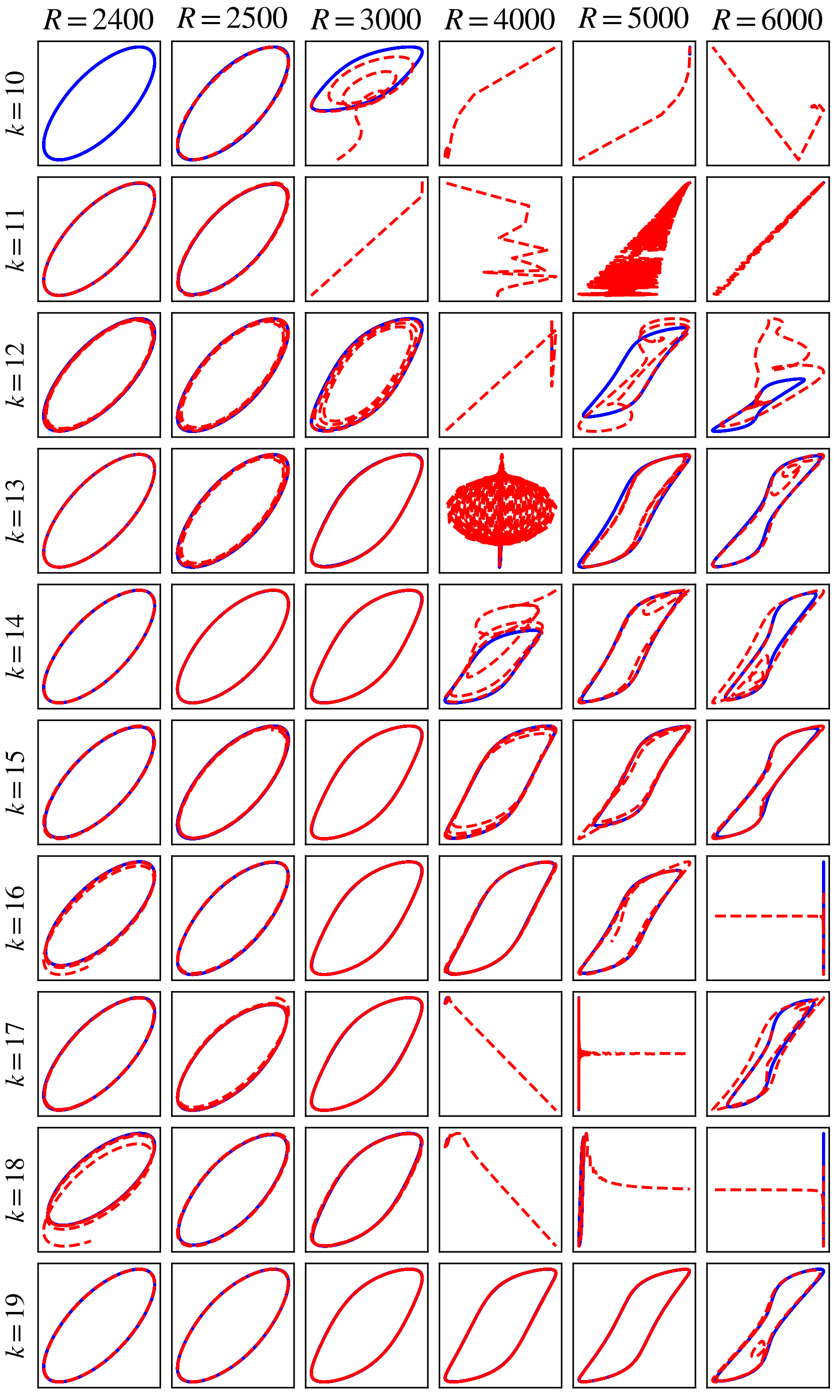


Figure 77: Comparison of integrations showing the $a - b$ plane (a on the horizontal and b on the vertical) for different sparsities and varying Rayleigh numbers with a second-order polynomial library. The training data is shown in blue and integration of the SINDy model in dashed red. Multiple revolutions in the $a - b$ plane can make some red dashed lines appear solid.

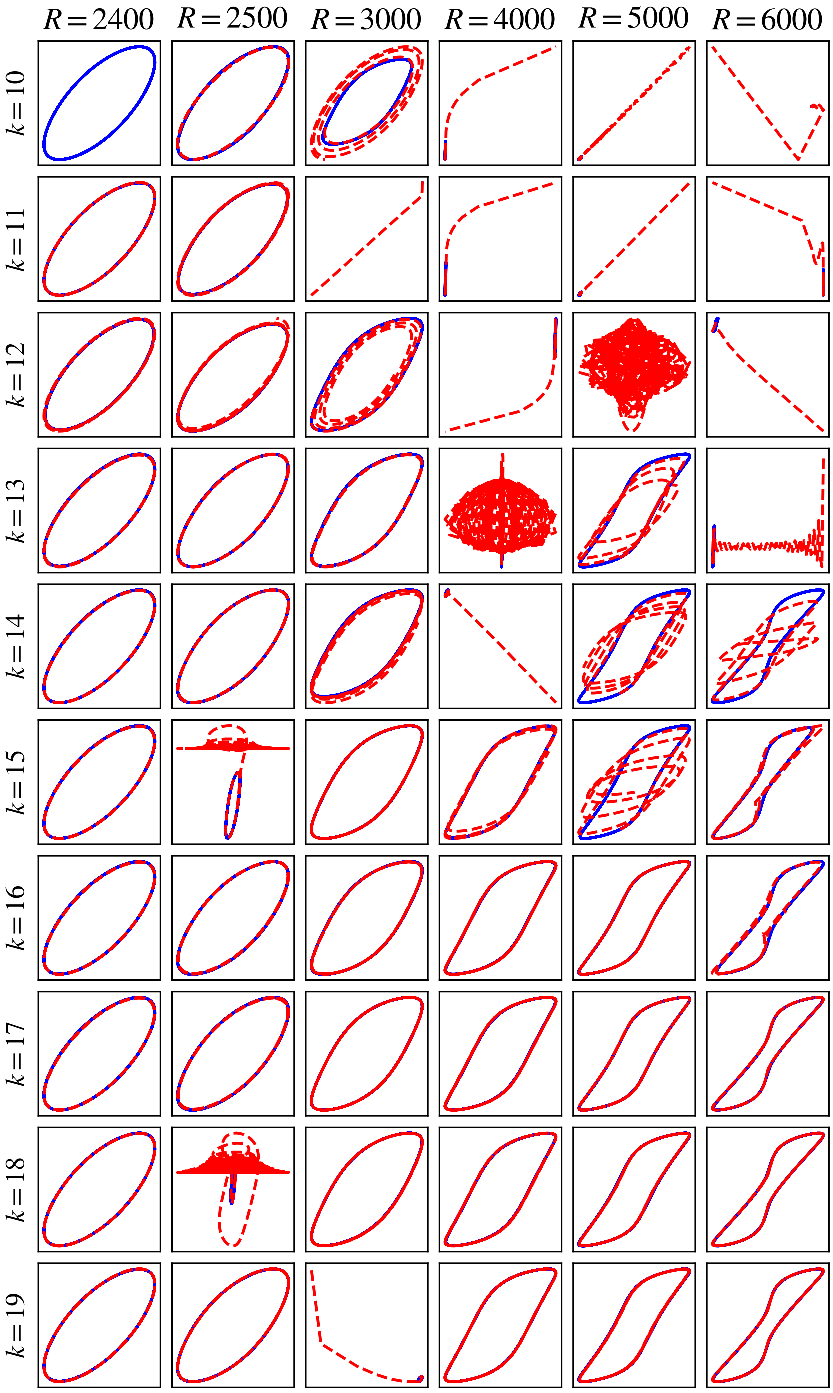


Figure 78: Comparison of integrations showing the $a-b$ plane (a on the horizontal and b on the vertical) for different sparsities and varying Rayleigh numbers with a third-order polynomial library. The training data is shown in blue and integration of the SINDy model in dashed red. Multiple revolutions in the $a-b$ plane can make some red dashed lines appear solid.

5.2.9 Closer analysis of select model behaviour

In this section we describe a“refitting” process to assess the suitability that a model derived at a fixed Rayleigh number can also be used to suitably describe the dynamics at other Rayleigh numbers, provided the coefficients are changed accordingly. To perform this process, we first take a model derived at a fixed R and fixed sparsity k (we refer to this as the first model). The first model identified at fixed R will have k non-zero coefficients corresponding to active terms in the feature library. We then perform a model fit at a different R with the same k, where the only allowed terms in the library are given by the non-zero terms identified in the first model. Consequently, only the coefficients of the identified original equations are allowed to change.

We first consider the refitting approach applied to the results of the second-order library, shown in Figure 79. Two different cases are plotted, the first where the model for $k = 20$ derived at $R = 6000$ is used to refit at all lower Rayleigh numbers. In this case, the high Rayleigh number model is capable of being refit accurately to all Rayleigh numbers. We similarly take the model derived at $k = 14$ and $R = 2400$ and refit this model to all higher Rayleigh numbers. In this case, the regression is not capable of fitting a model with only the active coefficients of the $R = 2400$, $k = 14$ model.

We similarly repeat the process described above with the third-order library, shown in Figure 80 and observe a similar pattern. The model derived at $R = 6000$ and $k = 17$ is capable of being refit to lower Rayleigh numbers, albeit less accurately than the second-order polynomial case. The model found at $k = 12$ and $R = 2400$ is, however, not capable of reproducing the dynamics at higher Rayleigh numbers.

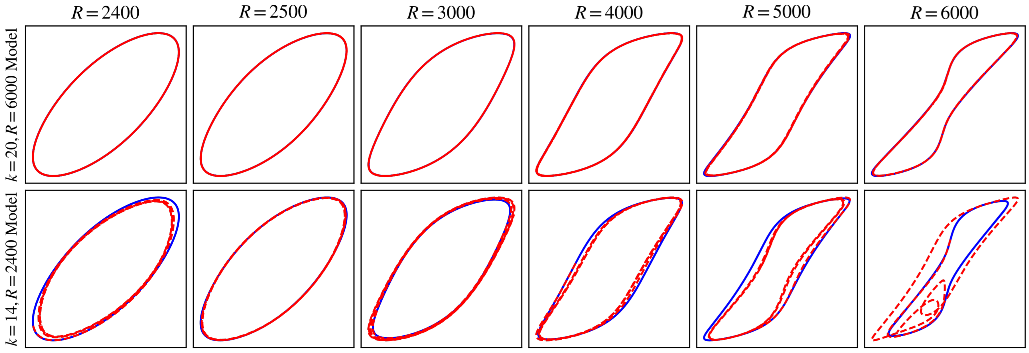


Figure 79: Refitting of models derived at $R = 6000$ and $R = 2400$ for the second-order polynomial library. Each row shows the integration of the refitted model from the specified Rayleigh number (in dashed red) against the reserved test data (in blue).

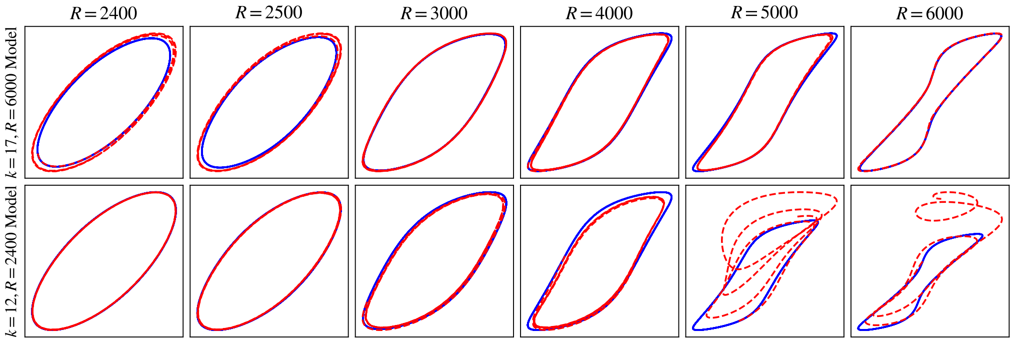


Figure 80: Refitting of models derived at $R = 6000$ and $R = 2400$ for the third-order polynomial library. Each row shows the integration of the refitted model from the specified Rayleigh number (in dashed red) against the reserved test data (in blue).

5.2.10 Extrapolation of a model to different Rayleigh numbers

As the results in the previous section suggest that models derived at $R = 6000$ are capable of being refitted to lower Rayleigh numbers, we can attempt to interpolate the solutions by either fitting the normalised Rayleigh number as an extra equation of motion such that $\dot{r} = 0$, or varying coefficients in the identified equations of motion. Parameterising SINDy models can then help reveal the bifurcation structure of the full PDE. The main challenge results from POD not identifying a common basis across different Rayleigh numbers, so we can expect that the form of the identified equations does not remain consistent at different Rayleigh numbers. The refitting process in Figure 79 was capable of refitting a model with the same active terms with different coefficients, however, when refitting coefficients do not appear to vary linearly with r. This is in contrast to the weakly nonlinear model, where the Rayleigh number drives convection through only the b term in the $\dot{a}$ equation. In our case as the basis decomposition varies, the Rayleigh number is bound to enter in multiple terms.

We first consider parameterising a model by introducing $\dot{r} = 0$ in the system of equations. An unconstrained fit with $k = 20$ for a quadratic library where $k = 20$ is chosen from the result that a model of this sparsity is required to fit to $R = 6000$. To train the model across different values of R, we construct a training data set from Dedalus runs for $R \in [2400, 2500, 2600, 2700, 3000, 3500, 4000]$. We restrict the range of R because Figure 68 shows that the identified POD basis modes change to a larger degree at higher Rayleigh number and in our tests the parameterised models performed poorly when including these data-sets. Identification of a model then yields the following set of ODEs

$$\begin{aligned}
\dot{a} &= -3.010b - 2.027ac + 2.307ae + 9.500bc + 1.172br + 8.693cd - 3.307de + 0.386dr,\\
\dot{b} &= 1.001a - 1.005b + 0.978ac,\\
\dot{c} &= -2.091c - 1.941ab,\\
\dot{d} &= -1.136a + 0.384b - 1.070ae,\\
\dot{e} &= -0.785ee + 2.028ad + 0.699ce - 0.156er,\\
\dot{r} &= 0.
\end{aligned}$$

At this point we can see the identification of three terms in the Rayleigh number associated to br, dr and er. This suggests that, as the POD basis varies between Rayleigh numbers, we also have variation in the amplitude of these modes. Comparisons of the model evaluated at different Rayleigh numbers is presented in Figure 81. This model parameterises trajectories in the range it is trained on well, growing to the expected amplitudes. It successfully predicts a bifurcation to no convection at $R = 2300$ which is expected numerically (ref. [8] lists the expected value as $R^{(o)} \approx 2306$). However, at $R = 6000$ the oscillations qualitatively match the POD time series, but the amplitude is incorrect. This undoubtedly arises from the difference in basis modes. At $R = 7000$, Dedalus simulations show steady convection which is not predicted by the identified model. A further increase of r still does not produce a bifurcation to steady convection.

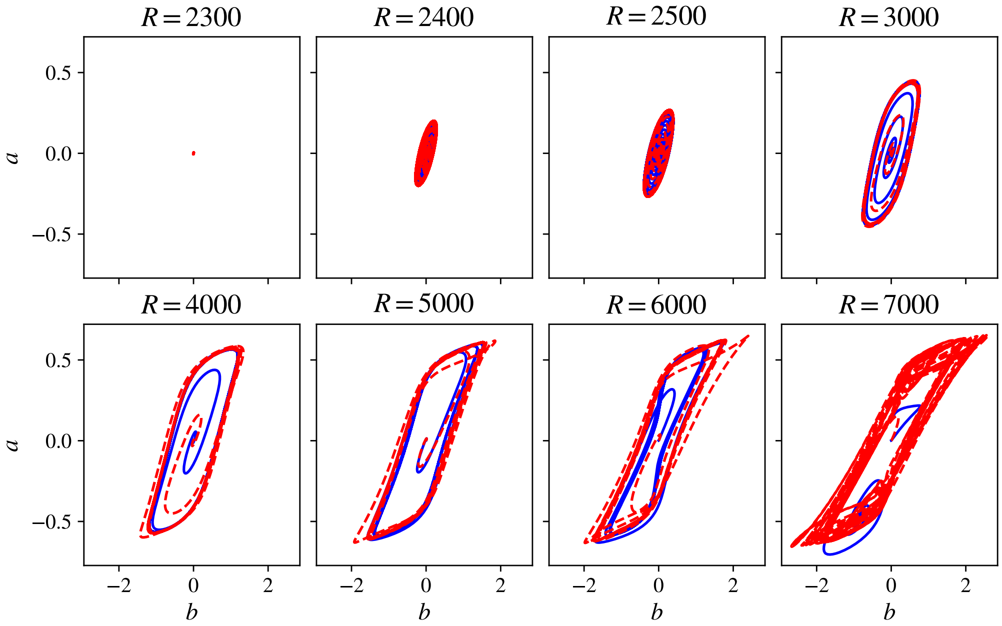


Figure 81: Comparison of the $a - b$ plane between the parameterised model (red) and the POD models (blue). The results are $R = 2300$ are steady and no POD mode is plotted here.

Despite challenges parameterising the models related to the basis decomposition, models derived at fixed Rayleigh number can still exhibit solution behaviour that they were not trained on. If we consider the model derived at $R = 6000$ with both constraints and a fixed sparsity of $k = 17$, we obtain the result

$$\begin{aligned}
\dot{a} &= 4.098b + 1.059d - 2.091de - 7.767bce, \\
\dot{b} &= 1.015a - 1.053b + 0.980ac, \\
\dot{c} &= -1.687c - 1.908ab - 0.579cdd, \\
\dot{d} &= -1.084a - 1.775be + 1.562ace, \\
\dot{e} &= -0.398e + 0.257ab + 1.234ad - 0.706abe.
\end{aligned}$$

If we assume as with the weakly nonlinear model that the Rayleigh number features on the coefficient of b in the $\dot{a}$ equation (denoted ξ_b), we can plot integrations of the model for different values of the coefficient shown in Figure 82. Each integration is given the initial condition 0.01 for all state variables, so can produce some transient behaviour. We can see that despite being trained at a fixed Rayleigh number, varying the coefficient of b still produces solutions which are stable at $\xi_b = -1$, with a transition to overstable oscillations that become progressively more nonlinear until an eventual bifurcation to steady convection between $\xi_b = 4.11$ and $\xi_b = 5$. The results suggest that, with a common basis it would be possible to parameterise the resulting models, though this remains for future work.

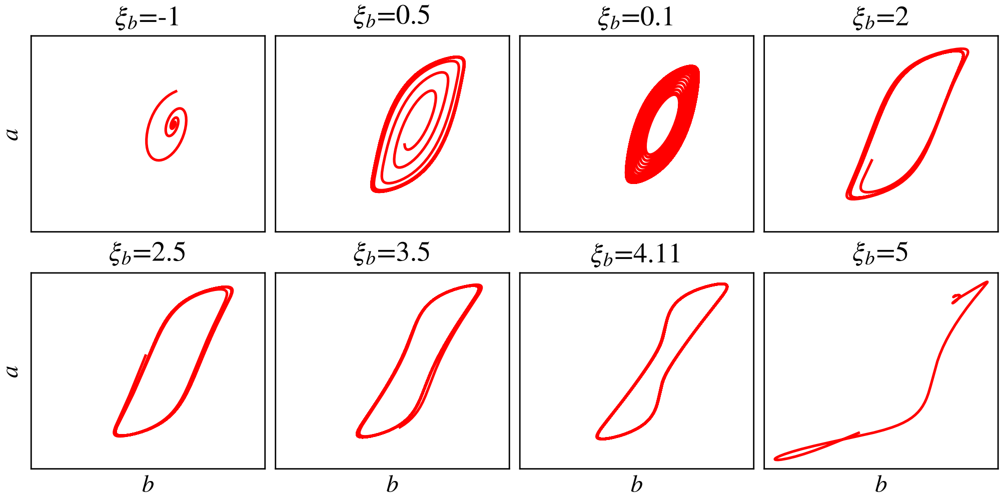


Figure 82: Integration of the constrained cubic model showing the $a - b$ plane while varying the coefficient magnitude for b in the $\dot{a}$ equation denoted ξ_b. As the coefficient is varied there is a transition from no convection (top left) to oscillatory convection which becomes progressively more nonlinear until an eventual transition to steady convection (bottom right).

5.3 Chaotic results

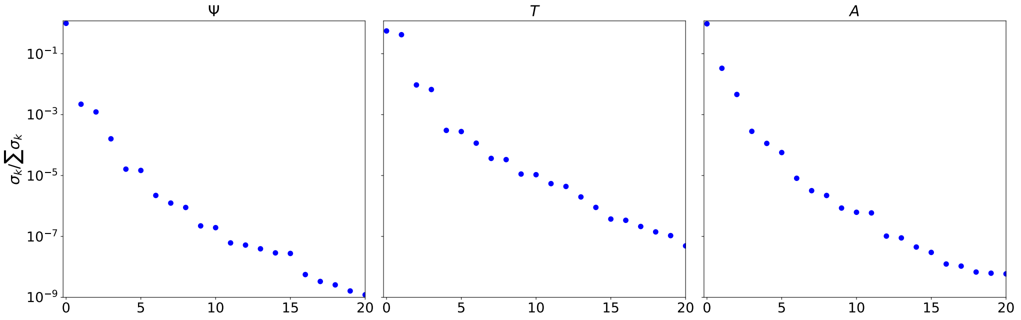
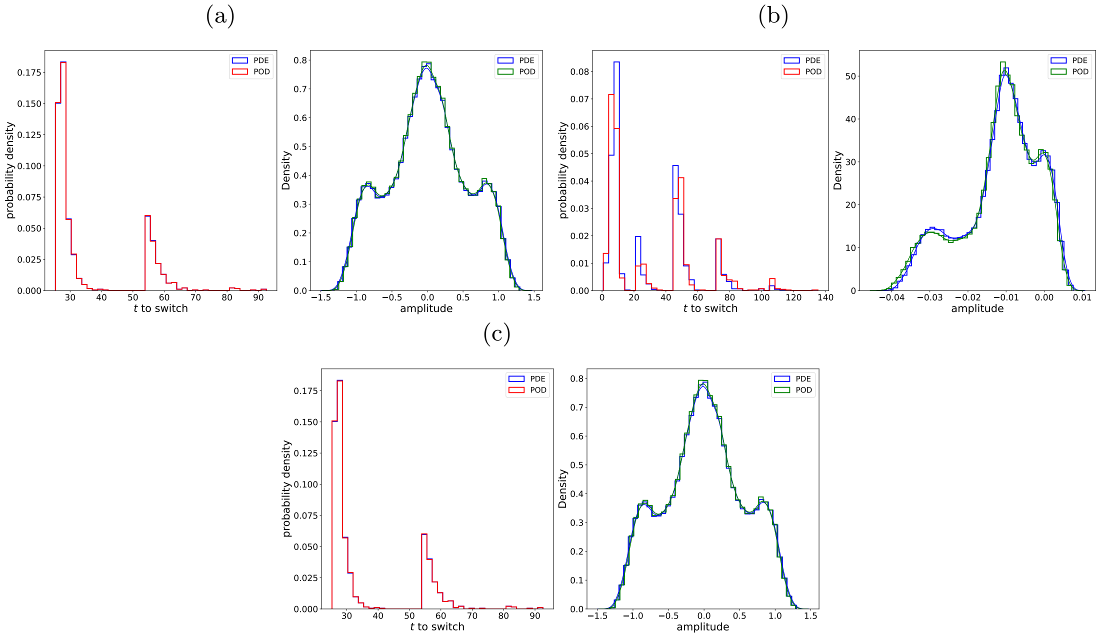


Figure 83: POD decomposition of ψ, T and A for the chaotic case.

We now study a single case where the convection is chaotic for the simulation parameters

$$R = 1088340, \quad Q = 85862, \quad \zeta = 0.8, \quad Pr = 1 \tag{5.34}$$

and $L_x = 0.16, \varpi = 1$ representing a vertical slot given in ref [25]. The training data is limited to 100 Lyapunov times with data sampled at a simulation time-step of 0.001. The singular values of the POD decomposition are shown in Figure 83. In this instance, we find that 4 modes in total are sufficient to reproduce the statistics of the flow. Figure 84 shows the original PDFs of the full order system against the truncated POD reproduction, as well as the corresponding switching times. The switching times here indicate the time taken for the listed variables to cross from positive to negative values, hence giving an indication of how long is spent in one lobe of the attractor. The PDFs for ψ and A exhibit strong symmetry, indicating that cells rotating clockwise or anti-clockwise are equally likely. For all fields we find that we only need one mode for Ψ and A, and two for T to successfully capture the steady statistics of the system. We justify construction of SINDy models from ψ_0, T_0, T_1 and A_0 based on singular-value distribution and reconstruction of the PDFs.


(a)

(b)

(c)

Figure 84: Reproduction of the full system statistics from the POD decomposition for a) ψ, b) T, and c) A. The left figure shows the switching times and the right shows the PDFs of the PDE and POD time series.

In the chaotic case, some care must be taken when evaluating model performance using integration of the resulting SINDy model. This is because, as the behaviour is chaotic, we expect that trajectories will inevitably diverge. For model selection we look for sparse models which are able to reproduce the PDFs of the full system and predict accurately over a large number of Lyapunov times. To assess the similarity of the PDFs, we use the multivariate KL divergence score given by

$$KL(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}) = \frac{1}{2}\left(\text{Tr}(\hat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - n + \ln\frac{|\hat{\mathbf{\Sigma}}|}{|\mathbf{\Sigma}|}\right)$$

where $\mathbf{\Sigma}$ is the variance-covariance matrix from the POD data, $\hat{\mathbf{\Sigma}}$ is the variance-covariance matrix of the integrated model and n the number of equations in the regression (4 here), as used by [138] (with definitions in §3.2). For clarity, we emphasise the covariance matrices are calculated on the state variables and not the derivatives. Integrating the models is more expensive to evaluate as the resulting model is integrated for a long time. We also calculate the number of Lyapunov times successfully predicted by the model (given in ref. [134]), defining the normalised error as

$$E(n) = \frac{||\boldsymbol{y}(n) - \hat{\boldsymbol{y}}(n)||}{\langle ||\boldsymbol{y}||^2\rangle^{1/2}} \tag{5.35}$$

where $\langle\cdot\rangle$ denotes the time average, $\boldsymbol{y}$ the observed data and $\hat{\boldsymbol{y}}$ the predicted data from the SINDy model. We then measure the number of predicted Lyapunov times before the normalized error exceeds 2. The chosen error threshold worked well for the particular series considered, however equation (5.35) is still an absolute measure of error and so the tolerance depends on the amplitude of the series. As a result, when the amplitude of $\boldsymbol{y}$ is small, the error can be large. We finally also use the corrected Akaike information criterion (AIC)

$$\text{AIC}_c = m\ln(\text{RSS}/m) + 2k + \frac{2(k+1)(k+2)}{m-k-2} \tag{5.36}$$

where the RSS is the residual sum of squared errors calculated from the predicted ($\dot{\boldsymbol{X}}_{\text{SINDy}}$) and true derivatives ($\dot{\boldsymbol{X}}_{\text{True}}$), k is the total number of non-zero coefficients and m is the number of samples. For a given length of training data, the training data is sampled a total of 30 times in different locations. The corrected AIC score is then computed for each identified model and averaged over the 30 different fits. This scoring method aims to balance model prediction and model sparsity and is not evaluated by integrating the identified model.

5.3.1 Tree-Parzen Estimators

An important aspect of successful system identification in SINDy is an appropriate choice of hyperparameters. Many aspects such as data sampling rate, data length, off trajectory dynamics and optimisation parameters can have a strong impact on the quality of the identified model, as has been seen so far. Tuning of these parameters is often performed using a standard grid-search approach. However, in most cases we would prefer a systematic approach to model selection which does not rely on fixing certain parameters. For example, practical advantages lie in using the target sparsity feature of MIOSR as this helps side-step normalisation issues by regressing each equation separately. However, this then opens an obvious question as to how to fix the sparsity of each equation. Further, if we wish to evaluate metrics of model performance which involve integration of the resulting model, grid search approaches rapidly become too expensive. One solution to this is the use of Tree-Structured Parzen Estimators (TPEs), developed by ref [81]. TPEs are a Bayesian optimisation framework which crucially make selection of new promising hyperparameters to test based on the history of performance of previously tested hyperparameters. This allows for optimisation of complex problems of many dimensions. A clear and comprehensive tutorial for TPEs is given by [168].

Several packages exist to tackle the use of hyper-parameter optimisation such as Optuna [117] and Hyperopt [94]. Both of these provide modular frameworks for tackling optimisation problems using a variety of different search algorithms. These methods aim to optimise what is known as the Expected Improvement (EI), defined as expectation that a function (the loss function) $f : \chi \to \mathbb{R}$ will exceed a threshold y^*

$$EI_{y^*} = \int_{-\infty}^{\infty} \max(y^* - y, 0) p_M(y|x) \mathrm{d}y \tag{5.37}$$

for some surrogate model of f named M. In this case f is expected to be a function which is expensive to evaluate, such as a loss function and leads us to construct a surrogate of the loss function. The expected improvement allows for a trade-off between exploration of the parameter-space and exploitation of previously known promising parameters. In this definition, x can be regarded as a single hyper-parameter or parameter being optimised, and y is the value of the user-defined loss function. To model this function, the distribution of $p_M(y|x)$ must also be modelled which can be done through, for example, kernel density estimation [168].

The aim of the optimisation is then to find an x^* that minimizes the loss function f, which can be costly to evaluate. TPEs do this by first modelling distributions $p(x|y)$ and $p(y)$ using the observation history. TPEs then maintain two distributions or surrogate models describing promising ($l(x)$) and non-promising ($g(x)$) hyper-parameters such that

$$p(x|y) = \begin{cases} l(x) & \text{if } y < y^* \\ g(x) & \text{if } y \geq y^* \end{cases} \tag{5.38}$$

where l is formed from the samples of x where the loss was less than y^* and correspondingly g is formed from the observations where the loss was greater than y^*. Finally, there remains freedom to choose y^* and this is chosen such that l is formed from some quantile γ, i.e., $p(y < y^*) = \gamma$. In the original paper by ref. [81], the authors showed that the EI is proportional to the ratio of probabilities of these two groups

$$EI_{y^*}(x) \propto \frac{l(x)}{g(x)} \tag{5.39}$$

so candidates which have high probability in l and low probability in g form valid candidates for selection. Through this ratio, the algorithm is able to select new hyperparameters for evaluation. Generally, hyperparameters are modelled independently, though the method can be modified for joint probabilities [168]. In the following, for each model selection method we use 10 trees initialised with different random seeds, each performing 300 trials with the optimisation run in parallel on one CPU. The allowable parameters are shown in Table 2, where we perform fits with MIOSR allowing the sparsity of each equation to vary independently from the others. The choices made here are quite restrictive and larger ranges could be used, but are based off of total sparsity fits for varying k. We also allow the total length of data to vary for a larger model space to be explored. In general though, if suitable models are not found the ranges can be extended. However, as we wish for sparse models the total sparsity for each equation is limited. Overall the largest expense when evaluating the TPEs in this case lie within the integrations of the model.

	α	No. Times	Eqn 1	Eqn 2	Eqn 3	Eqn 4
Range	[0.01, 0.1]	5,10, ... ,100	3,4,5,6	3,4,5,6	2,3,4,5,6	2,3,4

Table 2: Range of different parameters and hyper-parameter values used during TPE testing. These are chosen based on the general sparsity seen from the total sparsity method. The value of α varies logarithmically.

5.3.2 Overall performance

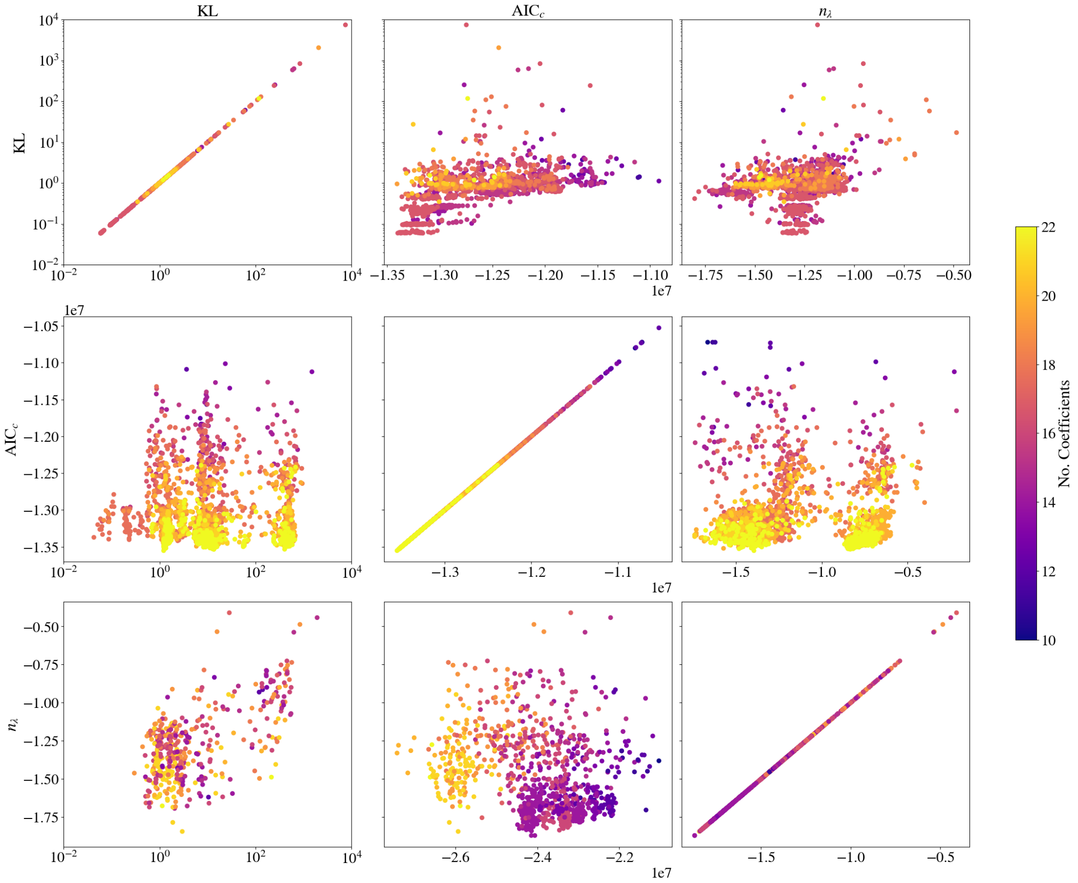


Figure 85: Comparison of the KL, AIC_c and n_λ scoring methods. The diagonal figures indicate the score being minimised on each row. For each row we also calculated the other scores, but these are not minimised.

We first consider the overall comparison of the different scoring methods when calculated using TPEs. The results of this process are shown in Figure 85. The diagonal figures (top left, middle and bottom right) show the scores being minimised for each row of the figure. For each score that is minimised, we also calculate the other corresponding model scores and plot these on the same row so comparisons can be made. As we are minimising all scores, we have plotted the negative of the predicted number of Lyapunov times. As a consequence, all of the best models evaluated by each metric are in bottom left of each plot. The colour gradient in each plot indicates the total number of non-zero coefficients for each model.

We now discuss the results of the scores that are minimised in each row. For the KL score, the lowest scoring models have around 17-19 non-zero coefficients, and models which have more non-zero coefficients perform worse. This is contrasted to the AIC_c score which produces the lowest scores for the lowest sparsity models (highest number of non-zero coefficients). For the Lyapunov scoring method, slightly sparser models are favoured compared to the KL scoring method, however while not visible, denser models also can produce low scoring models.

Comparison of the KL score to other scores calculated during minimisation along the row shows that models with a low AIC_c score also have low KL scores in this instance. For n_λ, we see that the lowest scoring KL models don't necessarily produce the best models for predicting n_λ. Minimisation of the AIC_c score in the middle row shows that only certain models with higher sparsity also produce low scoring KL models. Conversely, models with a large number of non-zero terms can produce models with a minimum n_λ. Finally for the minimisation of n_λ in the bottom row, we can see that the lowest scoring models minimised by n_λ do not typically produce low scoring KL models. As the assessment of model performance is only taken over a short time-frame, this may indicate that these models are not as good at capturing long term statistics. Compared to the AIC_c score, the sparser models which produce low n_λ do not produce minimal AIC_c scores.

The lack of correlation between the different metrics can be explained in the following way. The KL score measures the similarity of the resulting PDFs, and so has no measure of either the short-term predictive capability of the model or the quality of the derivative prediction given by the RSS contribution in the AIC_c score. Similarly for n_λ, only the short-term predictive ability of the model is assessed and, as will be discussed later, many of these models do not integrate favourably over longer time periods which negatively impacts their associated KL scores. For the AIC_c score, it is heavily influenced by the error between the predicted derivatives and the supplied derivatives, resulting in the AIC_c score favouring non-sparse models which typically do not perform well under the other scoring methods. Sparser models are favoured when integration of the resulting model is required, suggesting that sparser models could produce better predictive results.

5.3.3 Lyapunov Predictions

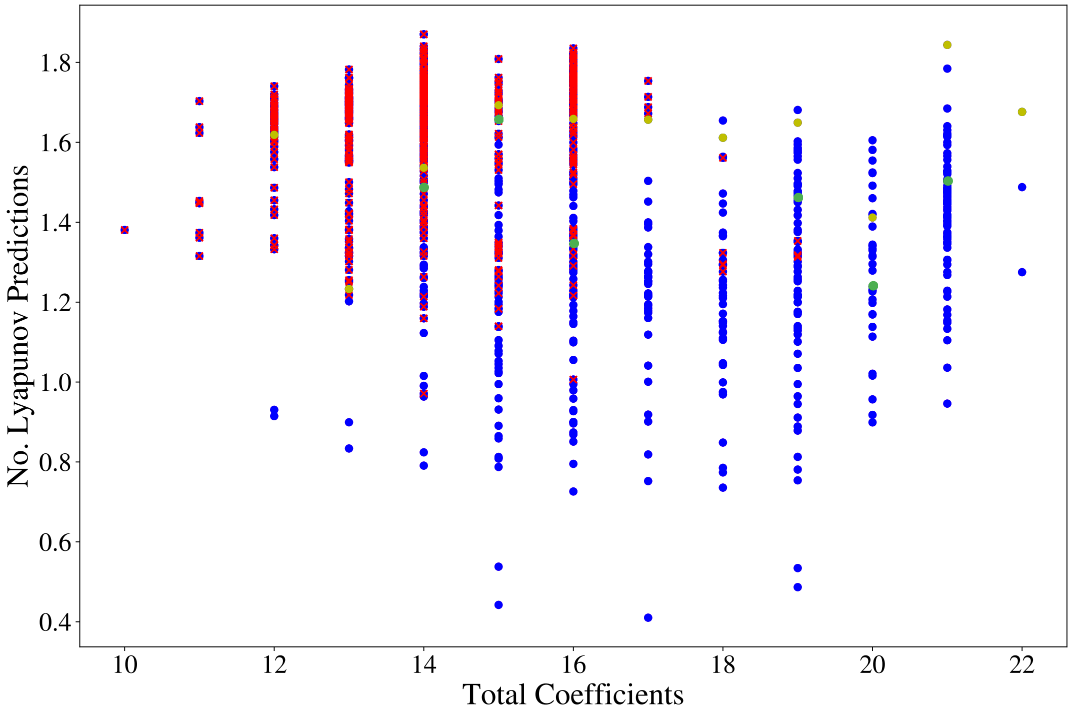
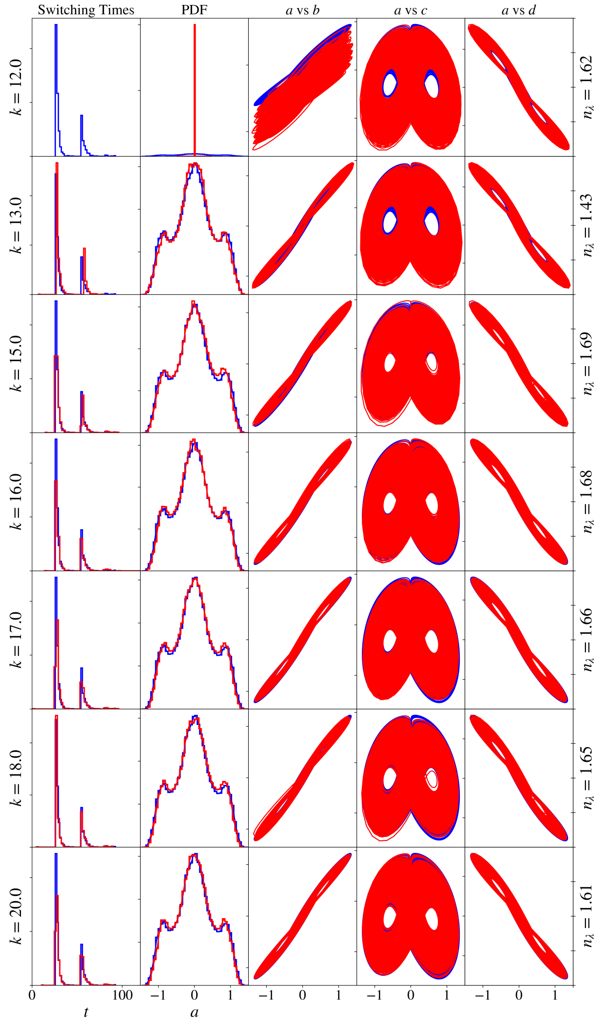


Figure 86: The Pareto front for models evaluating by the average number of Lyapunov times predicted over the test trajectory. Red crosses represent models which became unstable under longer integration. The green dots represent the models which obtain a minimum KL for a given number of coefficients and similarly the yellow obtain a minimum for the integrated AIC score.

We now consider models generated from maximising the number of Lyapunov times predicted by the models (remembering that the TPE minimises the negative of this). A more detailed view of the Pareto front is given in Figure 86. We plot all the identified models, integrating models for a longer time-period and marking models which fail to integrate with a red cross. In this case, we again limit the minimum time-step of the solver to 10^{-5}. We also mark the models which obtain a minimum KL score for each number of total coefficients in green, and similarly for the AIC_c score marked in yellow.

Results of the longer integrations suggest that many of the best performing sparse models do not integrate over longer time periods under the specified solver restrictions. After around 17 total coefficients, there are only a handful of models which no longer integrate for longer time periods. Above $k = 16$ the minimum AIC_c models also lie close to the maximum number of Lyapunov times predicted, but the KL scores do not.

We now compare the best performing models which also successfully integrate for a long time period in Figure 87. Several comparisons are made. Each row indicates the best performing model of a given sparsity, with the identified model being shown in the right-hand column. For each sparsity, we show a comparison of the switching times, PDF and on attractor dynamics of the training data (in blue) to the identified model (in red). We can see that beyond $k = 13$, the selected models all predict the dynamics well. As k is increased, many of the linear and quadratic terms remain consistent in the identified models, but the cubic terms tend to vary. For $k = 12$, the PDF does not compare well as when calculated the PDF integrations are started from a small initial condition. In this instance, SINDy has identified a model which does not become unstable with small initial conditions.



$\dot{a}$ = +0.7226a+0.4685d+1.1364ac
$\dot{b}$ = +0.7432a+0.5675d+0.2415ac+0.0063ddd
$\dot{c}$ = -0.1005c-0.0256aa-0.0064dd
$\dot{d}$ = -0.996a-0.8004d

$\dot{a}$ = +0.7227a+0.4685d+1.1361ac
$\dot{b}$ = +0.7183a+0.5429d+0.2876bc+0.0099abd
$\dot{c}$ = -0.1005c-0.0256aa-0.0064dd
$\dot{d}$ = -0.9983a-0.7986d+0.0074bbb

$\dot{a}$ = +0.7212a+0.4722d+0.8245ac+0.1813bc-0.0214bbb
$\dot{b}$ = +0.7191a+0.5451d+0.2651bc+0.012bbd
$\dot{c}$ = -0.1007c-0.0258aa-0.0063dd
$\dot{d}$ = -0.9961a-0.7969d-0.0033ddd

$\dot{a}$ = +0.7235a+0.4747d+0.8589ac-0.0966cd-0.0241bbb
$\dot{b}$ = +0.7219a+0.0632b+0.5945d+0.9214bc+0.4672cd+0.0064aad
$\dot{c}$ = -0.1005c-0.0256aa-0.0064dd
$\dot{d}$ = -0.996a-0.8004d

$\dot{a}$ = +0.7169a+0.4684d+0.774ac+0.2297bc-0.0228bbb
$\dot{b}$ = +0.7617a-0.0786b+0.5172d+0.2651ac+0.0485bc-0.0063aaa
$\dot{c}$ = -0.1006c-0.0257aa-0.0063dd
$\dot{d}$ = -0.9961a-0.7969d-0.0033ddd

$\dot{a}$ = +0.728a+0.4787d+0.9023ac-0.067cd-0.0224bbb
$\dot{b}$ = +0.7223a+0.0615b+0.5935d+0.9197bc+0.4656cd+0.0064aad
$\dot{c}$ = -0.0995c+0.0156ad-0.0161aa
$\dot{d}$ = -0.9958a+0.0426b-0.7611d+0.0434cd

$\dot{a}$ = +0.7124a+0.4623d+0.8162ac-0.0025ad-0.1586cd+0.0162abd
$\dot{b}$ = +0.764a-0.0726b+0.5266d+0.2739ac-0.006add
$\dot{c}$ = -0.0034a+0.0031b-0.1002c-0.0215aa-0.0154bb
$\dot{d}$ = -0.9971a+0.0482b-0.7579d+0.0437cd

Figure 87: A comparison of models along the Pareto front which predict, on average, a maximum number of Lyapunov times. Each row of the Figure shows a model of different sparsity with comparison of the switching times, PDF, on attractor dynamics and the resulting model shown in the right-column.

5.3.4 KL

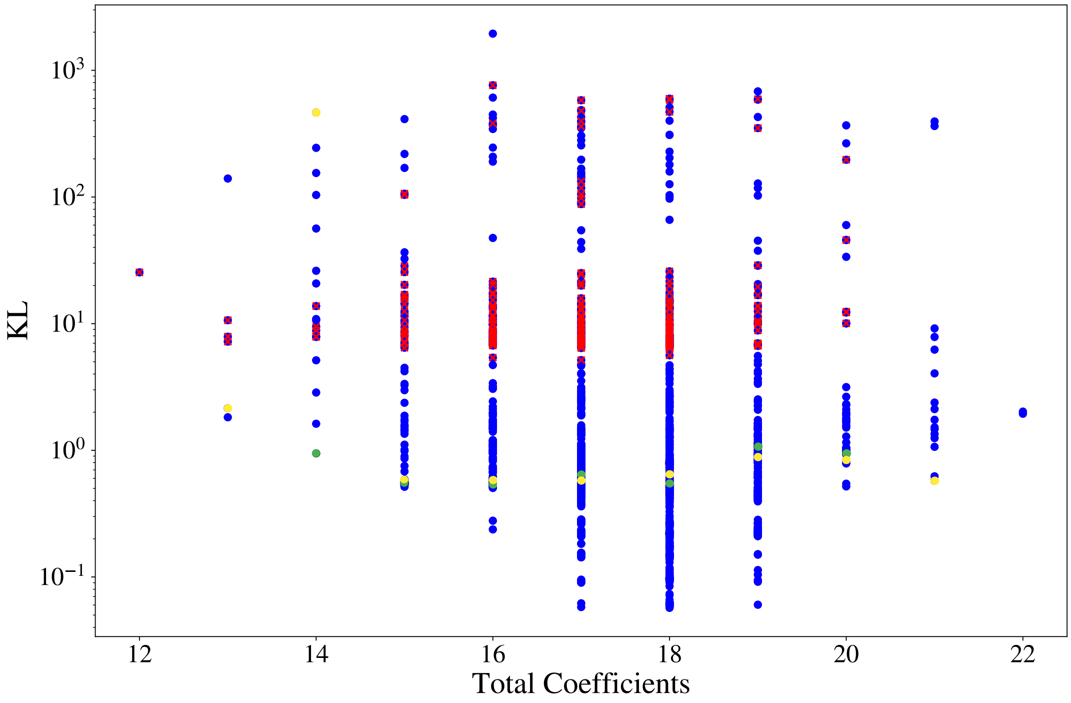
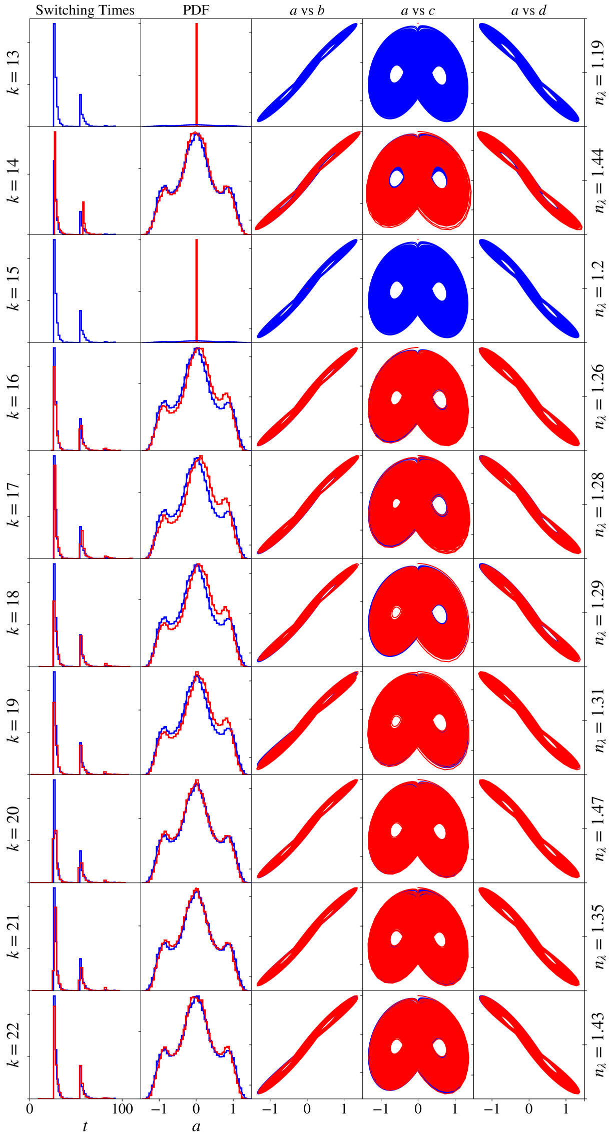


Figure 88: Scores for different models during the model selection process. The blue dots show all identified models, while the red crosses represents models which became unstable while finding the average number of predicted Lyapunov times. In this instance, yellow dots are models which have a maximum number of predicted Lyapunov times for a given number of coefficients, and green dots have a minimum AIC score.

We next compare the models generated from the trees which attempt to minimise the KL score. All models are shown in Figure 88, where this time we have marked models which become unstable at some point when calculating the n_λ score (again with the minimum solver time-step set to 10^{-5}). We can see that the KL score obtains a minimum value for around $k = 17 - 19$ coefficients. Also, KL scores which find a score below a given value also successfully integrate under the n_λ scoring method, even though this is not minimised against. Models with maximum number of Lyapunov times predicted do not seem to appear in conjunction with minimal KL models. We are, however, applying the KL score to distributions which are not Gaussian. The KL score in the form used here only compares the covariance of the distributions and so further minimisation of the score beyond a given point may not accurately capture non-Gaussian aspects of the distributions.

Figure 89 shows models with a minimal KL score along the Pareto front along with comparisons of the switching times, PDFs and on attractor dynamics. In the right hand y-label we list the corresponding n_λ which shows that the models selected through the KL score typically predict fewer n_λ than models obtained by optimising for n_λ. We can also see that for some sparsities, models are obtained which show a slight preference for one lobe of the attractor compared to the other, indicated by a skew in the PDF. Again the KL score is only comparing the covariance matrices and this difference will not be accounted for. In general, the models identified still produce reasonable predictions of the on attractor dynamics.



$\dot{a}$ = +0.723a+0.469d+1.1346ac
$\dot{b}$ = +0.7357a+0.5561d+0.3472ac
$\dot{c}$ = -0.0037a+0.0036b-0.0994c+0.0157ad-0.0158aa
$\dot{d}$ = -0.9961a-0.8005d

$\dot{a}$ = +0.7227a+0.4685d+1.1363ac
$\dot{b}$ = +0.7324a+0.5534d+0.7961bc+0.3637cd-0.0081aab
$\dot{c}$ = -0.1004c-0.0256aa-0.0064dd
$\dot{d}$ = -0.9983a-0.7986d+0.0074bbb

$\dot{a}$ = +0.6219a+0.3937d-0.6753cd+0.0213ddd
$\dot{b}$ = +0.7336a+0.5565d+0.7412bc+0.3462cd-0.0122bbb
$\dot{c}$ = -0.0004a-0.0993c+0.0154ad-0.0162aa
$\dot{d}$ = -0.996a-0.8004d

$\dot{a}$ = +0.7137a+0.4643d+0.7533ac-0.0023ad+0.2636bc-0.0128add
$\dot{b}$ = +0.7642a-0.0735b+0.5261d+0.2742ac-0.006add
$\dot{c}$ = -0.1004c-0.0256aa-0.0064dd
$\dot{d}$ = -0.996a-0.8004d

$\dot{a}$ = +0.7361a+0.4884d+0.9442ac+0.0027aa+0.0119ddd
$\dot{b}$ = +0.763a-0.0749b+0.5239d+0.2638ac-0.0086cd-0.0061add
$\dot{c}$ = -0.1005c-0.0257aa-0.0064dd
$\dot{d}$ = -0.9983a-0.7986d+0.0075bbb

$\dot{a}$ = +0.7132a+0.464d+0.8142ac-0.1496cd+0.0035bb-0.013add
$\dot{b}$ = +0.763a-0.0749b+0.5239d+0.2639ac-0.0086cd-0.006add
$\dot{c}$ = -0.1005c-0.0257aa-0.0064dd
$\dot{d}$ = -0.9983a-0.7986d+0.0075bbb

$\dot{a}$ = +0.7132a+0.464d+0.8144ac-0.1496cd+0.0035bb-0.013add
$\dot{b}$ = +0.721a+0.0654b+0.595d+0.933bc+0.4713cd-0.0079aab
$\dot{c}$ = -0.1006c-0.0257aa-0.0063dd
$\dot{d}$ = -0.9958a+0.0429b-0.7609d+0.0434cd

$\dot{a}$ = +0.7132a+0.464d+0.8141ac-0.1498cd+0.0035bb-0.013add
$\dot{b}$ = +0.7617a-0.0785b+0.5172d+0.2652ac+0.0484bc-0.0063aaa
$\dot{c}$ = -0.0036a+0.0034b-0.1003c+0.0098bd-0.0239aa
$\dot{d}$ = -0.9849a-0.7853d+0.0412cd

$\dot{a}$ = +0.7123a+0.4631d+0.8097ac-0.1552cd+0.002dd-0.0129add
$\dot{b}$ = +0.7643a-0.0711b+0.5256d+0.3086ac-0.006aaa
$\dot{c}$ = -0.0035a+0.0033b-0.1002c-0.0005ab-0.0213aa-0.0152bb
$\dot{d}$ = -0.9926a+0.0298b-0.7684d+0.0428cd

$\dot{a}$ = +0.7128a+0.4636d+0.8121ac-0.0027bd-0.1523cd-0.013add
$\dot{b}$ = +0.7227a+0.0592b+0.592d+0.9138bc+0.4611cd+0.0065aad
$\dot{c}$ = -0.0034a+0.003b-0.0994c-0.0527ab-0.0129ad-0.0014aa
$\dot{d}$ = -0.9942a+0.0362b-0.7648d+0.0431cd

Figure 89: A comparison of models along the Pareto front which predict minimal KL scores. Each row of the Figure shows a model of different sparsity with comparison of the switching times, PDF, on attractor dynamics and the resulting model shown in the right-column.

5.3.5 AIC_c

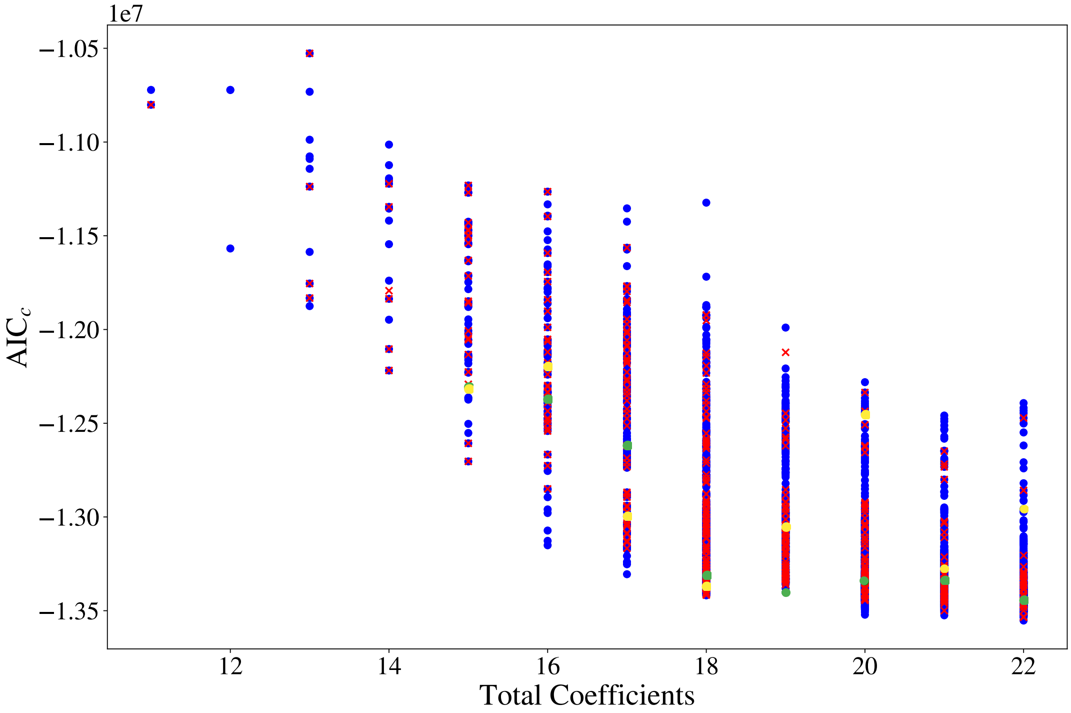


Figure 90: Scores for different models during the model selection process. The blue dots show all identified models, while the red crosses represents models which became unstable while finding the average number of predicted Lyapunov times. In this instance, yellow dots are models which have a maximum number of predicted Lyapunov times for a given number of coefficients, and green dots have a minimal KL score.

For the AIC_c, we show the minimum scoring models in Figure 90. In this instance, we again mark models which become unstable when calculated n_λ by a red cross. Observation of this indicates that low sparsity models with minimal AIC_c scores do not typically integrate successfully. As with many other examples considered until now, the AIC_c score does not show a clear increase in error again as k increases. There is an elbow in error around $k = 18$ which agrees with sparsities obtained by other methods. However, comparisons of the integrations of the models lying on the Pareto front shows that the majority of the models perform poorly (not shown). Many of them fail to reproduce both the PDFs and on attractor dynamics, suggesting that the evaluation of the derivatives in this case has not led to successful model identification. This observation typically agrees with previous results where we attempted to use the AIC_c method. Strangely the score has seen success in other SINDy applications such as ref. [163] and [105], so we suggest that the poor performance here must be related to the basis only approximately describing the dynamics. Comparisons of the predicted derivatives to the supplied derivatives show that the largest errors come at the largest amplitudes, which will inevitably have the largest contributions in RSS error.

5.3.6 Comparisons to the weakly nonlinear model

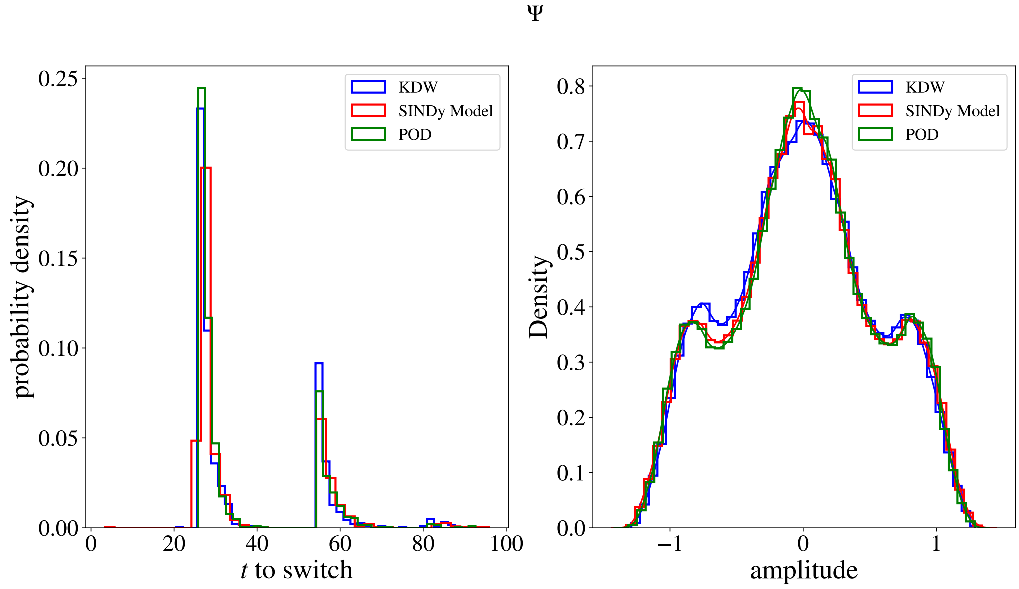


Figure 91: Comparisons between the POD data (green), SINDy model (red) and Knobloch model named KDW (blue). The left figure shows histograms of the switching times, showing how long trajectories typically take to move between a and $-a$. The right figure shows PDFs of the different variables.

While many models have been generated which can be argued as suitable, we consider one model here. We compare the model from the Pareto front for minimising n_λ with 15 total coefficients

$$\dot{a} = 0.721a + 0.472d + 0.824ac + 0.181bc - 0.021bbb, \tag{5.40}$$

$$\dot{b} = 0.719a + 0.545d + 0.265bc + 0.012bbd, \tag{5.41}$$

$$\dot{c} = -0.101c - 0.026aa - 0.006dd, \tag{5.42}$$

$$\dot{d} = -0.996a - 0.797d - 0.003ddd \tag{5.43}$$

to the performance of the weakly nonlinear model. This model obeys the symmetries

$$a \to -a, \quad b \to -b, \quad c \to c, \quad d \to -d \tag{5.44}$$

similar to the weakly nonlinear model, without prior enforcement. This results from the underlying data exhibiting this symmetry. Comparisons of the reproduced switching times and PDFs are shown in Figure 91 showing a favourable comparison between the SINDy model and the POD data. We also compare a single point of integration on the test set of data against the POD and Knobloch model in Figure 92. While the average number of Lyapunov times predicted by most models identified is ≤ 2, we can see that in this instance at least the identified model predicts well for more than 2 Lyapunov times. This is because the error criterion tends to become large when the amplitude of the time series considered is small. As mentioned, because it is a relative error measure, errors can become large close to zero.

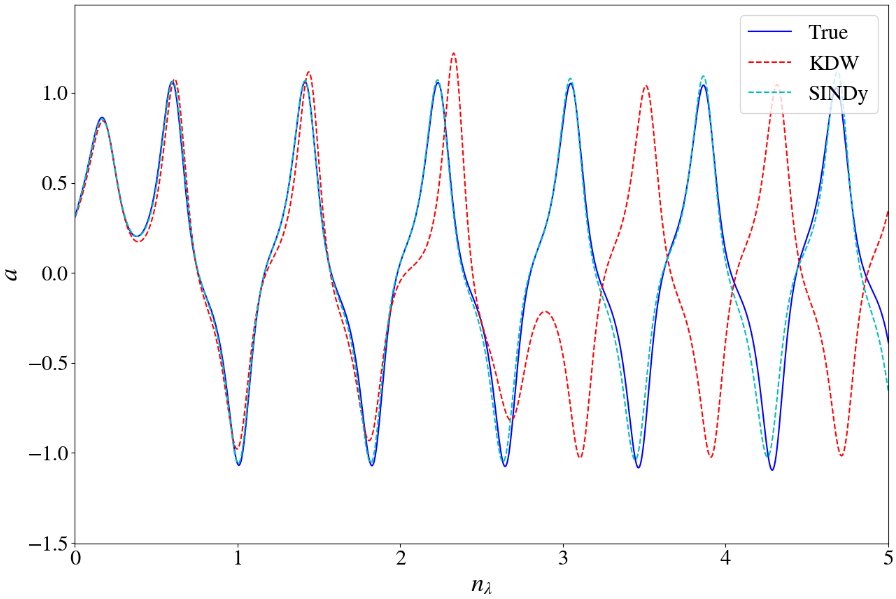


Figure 92: Comparison of the integrated trajectories of the POD data (blue), SINDy model (dashed blue) and the Knobloch model named KDW (dashed red).

5.4 Conclusions

In this chapter we applied SINDy to time series from a POD decomposition of PDE data from simulations of the magnetoconvection equations given in § 4.1. This represents an interesting test case for SINDy as close to the onset of overstable oscillations there is a known solution given by weakly nonlinear theory. Further from the onset of overstable oscillations, the POD modes deviate from Fourier modes and we then require SINDy to construct different models. We first discussed issues surrounding the selection of sparse models using STLSQ with SINDy. In all cases examined, using the AIC score never favoured a sparse model. For STLSQ, this makes selecting a viable threshold challenging as sparse models give poor model scores, even in cases where we know an analytical solution exists with smaller k (number of non-zero coefficients). As a result, we instead use MIOSR as this allows the total number of non-zero coefficients to be controlled explicitly.

With MIOSR we considered the MSQE of model fits at each Rayleigh number using no constraints, symmetry constraints, diffusive constraints or both symmetry and diffusive constraints. We also calculated the MSQE for both a second-order and third-order polynomial library. Generally for low Rayleigh number the second-order library was capable of producing low MSQE models with few terms, and at higher Rayleigh number the cubic library produced lower MSQE models with fewer terms. In all cases, the constraints made a minor difference in MSQE but offer the advantage that models are guaranteed to obey the constraints and also make the regression faster. This is because inclusion of constraints reduces the number of possible features which can appear in the resulting equations and therefore reduces the required search time [160].

Integration of models of different sparsities under both constraints showed that for higher Rayleigh numbers, more terms are required to reproduce the on-attractor dynamics agreeing with the results suggested by the MSQE Pareto curves. Finally, fixing models to a specific k and parameterising the models showed some promise for capturing the bifurcation structure, but the results show that to do so successfully would require a common POD basis. This could either be achieved by take modes found at a fixed Rayleigh number and projecting them to other Rayleigh numbers, or performing POD so that a basis mode is found which best represents the data at all Rayleigh numbers. This remains for future work.

We also presented an approach for systematic model selection using TPEs and 3 different metrics. These metrics measured either the short-term predictive capability of identified models, the graphical fit of models balanced against their sparsity and the similarity of the long-term distributions. By using model selection metrics which require integration (KL and number of predicted Lyapunov times) we can identify suitably sparse models which compare favourably to the training data in the chaotic case. We also see that one resulting model is capable of reproducing the dynamics with one fewer mode that the weakly nonlinear model for fixed parameters.

For future tokamak applications, simulations of type I ELMs rely on the reduced-MHD single fluid equations which have parallels to the magnetoconvection problem discussed in [80]. Applications of the methods described here to simulation data from JOREK or BOUT++ could provide low-dimensional models describing ELMs. Computationally, these problems are very difficult to simulate as both the short timescales of the ELM crash and the long inter-ELM periods must be resolved [131]. Low-dimensional models provide a much less computationally demanding means of possibly exploring this type of behaviour. One aim of this project originally involved application of SINDy to BOUT++ ELM data, but at the time of study this code was not capable of simulated repetitive ELMs and so remains for future work.

6 A data assimilation approach with the ensemble Kalman filter

In this section we introduce the theory behind the Kalman and ensemble Kalman filters which are data assimilation approaches as an alternative approach for fitting the ANAC and ANAET models to experimental signals. A data-assimilation approach means we can take predictions from a numerical model and observations of a system and attempt to find a corrected state estimation. In this thesis, the aim is to use the ANAC or ANAET models as numerical models which can then make predictions of the magnetic field or temperature. By using a data assimilation approach, we can then take experimental observations from diagnostics on the tokamak and attempt to reconcile these with numerical predictions from symmetry-based equivariant bifurcation theory models.

We will first introduce the theory behind the ensemble Kalman filter and why this method has been chosen. We will then discuss several relevant modifications to this method which allow for the method to be more robust and account for different modelling errors. We then discuss several results of fitting the ANAC and ANAET models in progressively more challenging scenarios. Eventually the ANAET model is fitted in a synthetic case which closely resembles experiment.

6.1 Ensemble Kalman Filter Theory

In this section we describe how to construct a basic Kalman filter [3], where the notation will largely be kept in line with the description given by [24] as opposed to the previous work by CCFE student Luca Spinicci [115]. We start with a linear Kalman filter as understanding the ensemble Kalman filter follows directly as an ensemble average of the linear Kalman filter. The notation introduced here also differs to some extent from the EnKF implementation discussed later and these will be discussed.

6.1.1 The Gaussian Distribution

Before discussing the Kalman filter, we first define the Gaussian distribution which is relevant to the Kalman filter. In 1D, this is characterised by mean μ and variance σ^2 where σ is called the standard deviation of the distribution. We can write this as

$$f(\psi) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\psi-\mu)^2}{2\sigma^2}\right) \tag{6.1}$$

giving the continuous probability density function. The mean of the distribution simply states where the distribution is centred, and the variance gives some measure of the spread of the samples around that mean. For a Gaussian distribution, $\mu \pm \sigma$ spans approximately 68% of the samples, $\mu \pm 2\sigma$ spans 95% of the samples and so on. The distribution function which represents the probability that an observation of Ψ can take a value less than or equal to ψ is defined

$$F(\psi) = \int_{-\infty}^{\psi} f(\psi) d\psi. \tag{6.2}$$

Most relevant for our discussion will be considering the statistics of random fields of variables $\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^n$ where $\boldsymbol{x} \in \mathbb{R}^m$. In our current definitions, we now have a covariance matrix which describes the variation of each variable with the other. We can write the multivariate Gaussian distribution as

$$f(\boldsymbol{\psi}) = \frac{1}{(2\pi)^{n/2}(\det\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{\psi}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\psi}-\boldsymbol{\mu})\right) \tag{6.3}$$

If we have $i = 1, 2, \ldots, N$ statistically independent samples, labeled $\boldsymbol{\psi}^i$ then the sample mean is defined

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\psi}] = (\mathbb{E}[\psi_1], \mathbb{E}[\psi_2], \ldots, \mathbb{E}[\psi_n])^T \tag{6.4}$$

where

$$\mathbb{E}[\psi_1] = \frac{1}{N}\sum_{i=1}^{N} \psi_1^i(\boldsymbol{x}) \tag{6.5}$$

represents the sample mean of each component of the random vector. The covariance between two different locations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ of the random fields is defined

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{x}_1, \boldsymbol{x}_2) &= \mathbb{E}[(\boldsymbol{\psi}(\boldsymbol{x}_1) - \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{x}_1)]) \otimes (\boldsymbol{\psi}(\boldsymbol{x_2}) - \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{x}_2)])], & (6.6)\\ &= \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{\psi}^i(\boldsymbol{x}_1) - \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{x}_1)]) \otimes (\boldsymbol{\psi}^i(\boldsymbol{x_2}) - \mathbb{E}[\boldsymbol{\psi}(\boldsymbol{x}_2)]) & (6.7) \end{aligned}$$

where $\otimes$ represents the outer product of two vectors.

6.1.2 The Kalman Filter

We consider a dynamical system where we have, for example, a numerical forecast of the future represented by $\boldsymbol{v}_j^f \in \mathbb{R}^n$ at time j and a set of experimental measurements denoted $\boldsymbol{y}_j \in \mathbb{R}^m$. However, we are aware of the fact that one or both of these estimates may be incorrect. Given these two estimates, we then wish to construct an improved approximation of the true state $(\boldsymbol{v}_j^t)$. We first start by writing, in the linear discrete case, a set of equations representing our system

$$\begin{aligned} \boldsymbol{v}_{j+1}^f &= \boldsymbol{M}\boldsymbol{v}_j^t + \boldsymbol{\xi}_j, & (6.8)\\ \boldsymbol{y}_{j+1} &= \boldsymbol{H}\boldsymbol{v}_{j+1}^t + \boldsymbol{\eta}_{j+1} & (6.9) \end{aligned}$$

where $\boldsymbol{\xi}_j$ represents Gaussian additive noise in the forecast and $\boldsymbol{\eta}$ is measurement noise. In each case the true-state evolved $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ and the observation operator $\boldsymbol{H} \in \mathbb{R}^{m\times n}$. We now wish to combine $\boldsymbol{v}_{j+1}^f$ and $\boldsymbol{y}_{j+1}$ in some way to provide some estimate of $\boldsymbol{v}_{j+1}^t$ which we denote $\boldsymbol{v}_{j+1}^a$. To estimate this problem, we must make some assumptions about the properties of the noise and errors in our system

$$\begin{aligned} \overline{\boldsymbol{\xi}} &= \boldsymbol{0}, & \overline{\boldsymbol{\xi}\boldsymbol{\xi}^T} &= \boldsymbol{Q}, & (6.10)\\ \overline{\boldsymbol{\eta}} &= \boldsymbol{0}, & \overline{\boldsymbol{\eta}\boldsymbol{\eta}^T} &= \boldsymbol{\Gamma}, & (6.11)\\ \overline{\boldsymbol{\xi}\boldsymbol{\eta}^T} &= \boldsymbol{0} & & & (6.12) \end{aligned}$$

where $\boldsymbol{Q}$ is process noise matrix and $\boldsymbol{\Gamma}$ is the measurement error covariance matrix. We assume that the noise for the state forecast and measurement are both Gaussian with zero mean and that there is no cross-correlation between these errors. We further explicitly define the forecast and analysis covariance matrices as

$$\boldsymbol{C}^f = \mathbb{E}[(\boldsymbol{v}^t - \boldsymbol{v}^f) \otimes (\boldsymbol{v}^t - \boldsymbol{v}^f)] \tag{6.13}$$

and

$$\boldsymbol{C}^a = \mathbb{E}[(\boldsymbol{v}^t - \boldsymbol{v}^a) \otimes (\boldsymbol{v}^t - \boldsymbol{v}^a)]. \tag{6.14}$$

These are also assumed to have Gaussian distributions.

6.1.3 The Bayesian formulation

In this section we describe the Bayesian filter problem, the derivation of which underpins the use of Bayes' formula in the ensemble Kalman filter. Understanding this derivation is beneficial in understanding some of the limitations of EnKF. In the following discussion we follow the outline by [155].

We are interested in solving the assimilation problem of combining a prediction and observation at step L given some previous set $V = (\boldsymbol{v}_0^T, \boldsymbol{v}_1^T, \dots \boldsymbol{v}_L^T)$ of predictions of the system state where $\boldsymbol{v}_l$ may be augmented to include parameters or other uncertain quantities. V is then essentially the vector or matrix which includes all uncertain quantities in the inference problem. We also define the collection of measurements by $\mathcal{Y} = \mathcal{H}(V) + \epsilon$ where ϵ describes the measurement error. By using a Bayesian approach, we can combine the state prediction and measurements to create an improved estimate of the state.

We start with the prior denoted

$$f(V)$$

where $f(\cdot)$ denotes the pdf of a random variable in this instance. This represents our previous belief in the probability of some values of V. Given this knowledge, we introduce measurements through the likelihood written as

$$f(\mathcal{Y}|V)$$

where we note this is a function of V alone as the measurements $\mathcal{Y}$ are fixed for some particular draw. We will then need a way of estimating the likelihood which is found through an expression in the measurement error $f(\mathcal{Y}|V) = f(\epsilon) = f(\mathcal{Y} - \mathcal{H}(V))$. Given assumed knowledge of $f(V)$ and $f(\mathcal{Y}|V)$, we use Bayes' formula to find an expression for $f(V|\mathcal{Y})$

$$f(V|\mathcal{Y}) = \frac{f(\mathcal{Y}|V)f(V)}{f(\mathcal{Y})}. \tag{6.15}$$

The denominator is a normalisation ensuring $\int f(V|\mathcal{Y})\mathrm{d}V = 1$. Currently, this description includes all available measurements and state up to the assimilation window L. We now make an important simplification by assuming that we can describe this system as a first-order Markov process meaning that only measurements and states at a previous time-step are required. Mathematically we can write this as

$$f(\boldsymbol{v}_L|\boldsymbol{v}_{l-1}, \boldsymbol{v}_{l-2}, \dots, \boldsymbol{v}_0) = f(\boldsymbol{v}_L|\boldsymbol{v}_{l-1}) \tag{6.16}$$

implying the future depends only on the present state. We can then write the prior as the product of independent probability density functions

$$\begin{aligned} f(V) &= f(\boldsymbol{v}_0)f(\boldsymbol{v}_1|\boldsymbol{v}_0)\dots f(\boldsymbol{v}_L|\boldsymbol{v}_{l-1}), \\ &= f(\boldsymbol{v}_0)\prod_{l=1}^{L} f(\boldsymbol{v}_l|\boldsymbol{v}_{l-1}). \end{aligned} \tag{6.17}$$

Further, given the assumption of uncorrelated measurement error and the expression of the likelihood in terms of measurement error, this allows us to write the likelihood as the product of independent probabilities

$$f(\mathcal{Y}|V) = \prod_{l=1}^{L} f(\boldsymbol{y}_l|\boldsymbol{v}_l). \tag{6.18}$$

Using Equations 6.17 and 6.18, we can write the general Bayesian formula Eqn 6.15 as

$$f(V|\mathcal{Y}) = \frac{\prod_{l=1}^{L} f(\boldsymbol{y}_l|\boldsymbol{v}_l) f(\boldsymbol{v}_l|\boldsymbol{v}_{l-1}) f(\boldsymbol{v}_0)}{f(\boldsymbol{y}_L)}. \tag{6.19}$$

This now describes a recursive problem where the solution at l can be expressed in terms of the prior at $l-1$. Explicitly

$$\begin{aligned} f(\boldsymbol{v}_1, \boldsymbol{v}_0|\boldsymbol{y}_1) &= \frac{f(\boldsymbol{y}_1|\boldsymbol{v}_1) f(\boldsymbol{v}_1|\boldsymbol{v}_0) f(\boldsymbol{v}_0)}{f(\boldsymbol{y}_1)}, & (6.20) \\ f(\boldsymbol{v}_2, \boldsymbol{v}_1, \boldsymbol{v}_0|\boldsymbol{y}_2, \boldsymbol{y}_1) &= \frac{f(\boldsymbol{y}_2|\boldsymbol{v}_2) f(\boldsymbol{v}_2|\boldsymbol{v}_1) f(\boldsymbol{v}_1, \boldsymbol{v}_0|\boldsymbol{y}_1)}{f(\boldsymbol{y}_2)}, & \\ \vdots \quad &= \quad \vdots & \\ f(V|\mathcal{D}) &= \frac{f(\boldsymbol{y}_L|\boldsymbol{v}_L) f(\boldsymbol{v}_L|\boldsymbol{v}_{L-1}) f(\boldsymbol{v}_{L-1}, \boldsymbol{v}_{L-2}, \dots \boldsymbol{v}_0|\boldsymbol{y}_{L-1}, \boldsymbol{y}_{L-2}, \dots \boldsymbol{y}_1)}{f(\boldsymbol{y}_L)}. & (6.21) \end{aligned}$$

In Equation 6.20, $f(\boldsymbol{v}_1|\boldsymbol{v}_0)$ represents the integration of $\boldsymbol{v}_0$ to give the pdf of $\boldsymbol{v}_1$ given $\boldsymbol{v}_0$. $f(\boldsymbol{v}_0)$ represents our prior belief in the initial condition $\boldsymbol{v}_0$ and finally $f(\boldsymbol{y}_1|\boldsymbol{v}_1)$ includes information on the measurements. This allows us to estimate the posterior represented by the joint PDF $f(\boldsymbol{v}_1, \boldsymbol{v}_0|\boldsymbol{y}_1)$ conditioned by measurements $\boldsymbol{y}_1$. At the next recursion step, this becomes the prior in our formula. This solution is known as the smoother solution to the Bayesian filtering problem [155].

Finally, we can integrate over the previous states and can represent the problem recursively in terms of marginal pdfs

$$f(\boldsymbol{v}_1, \boldsymbol{v}_0|\boldsymbol{y}_1) = \frac{f(\boldsymbol{y}_1|\boldsymbol{v}_1) \int f(\boldsymbol{v}_1|\boldsymbol{v}_0) f(\boldsymbol{v}_0) \mathrm{d}\boldsymbol{v}_0}{f(\boldsymbol{y}_1)} = \frac{f(\boldsymbol{y}_1|\boldsymbol{v}_1) f(\boldsymbol{v}_1)}{f(\boldsymbol{y}_1)}, \tag{6.22}$$

which allows us to derive the sequential formula

$$f(\boldsymbol{v}_L|\boldsymbol{y}_L) = \frac{f(\boldsymbol{y}_L|\boldsymbol{v}_L)}{f(\boldsymbol{y}_L)} f(\boldsymbol{v}_L) \tag{6.23}$$

which is exactly Bayes' formula. Our update scheme now only depends the results at the assimilation window L, and all previous states are implicitly held within $\boldsymbol{v}_L$. This relation will be used in the majority of EnKF implementations.

6.1.4 The Maximum "a Posterior" Estimate

Bayes' formula is a widely used and powerful expression which allows us to claim something about $f(\boldsymbol{v}_L|\boldsymbol{y}_L)$ given that we know about the right-hand side of Equation 6.23. Naturally we wish to have the most probable prediction given a particular observation and we can achieve this by maximising $f(\boldsymbol{v}_L|\boldsymbol{y}_L)$. This means we need only maximise the numerator of the right-hand side of Equation 6.23. To do this, we need to make some assumptions on $f(\boldsymbol{y}_L|\boldsymbol{v}_L)$ and $f(\boldsymbol{y}_L)$.

To maximise Equation 6.23 we consider the maximum "a posterior" estimate (MAP) as the solution which maximises

$$\boldsymbol{v}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{v}} (f(\boldsymbol{v}|\boldsymbol{y})) \tag{6.24}$$

for a given assimilation window. The can be written as an equivalent minimisation of

$$\boldsymbol{v}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{v}} \mathcal{J}(\boldsymbol{v}) \tag{6.25}$$

where $\mathcal{J}$ is the cost function in the expression

$$f(\boldsymbol{v}|\boldsymbol{y}) \propto \exp(-\mathcal{J}(\boldsymbol{v})). \tag{6.26}$$

We already stated our assumption that the errors in our formulation will be normally distributed, so we can write these as multivariate Gaussian distributions

$$f(\boldsymbol{v}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{v}^f)^T(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}-\boldsymbol{v}^f)\right) \tag{6.27}$$

and

$$f(\boldsymbol{y}|\boldsymbol{v}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{v}-\boldsymbol{y})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{v}-\boldsymbol{y})\right) \tag{6.28}$$

noting that the above expression of the likelihood corresponds to direct measurements which is true in our case. Thus the cost function is simply

$$\mathcal{J}(\boldsymbol{v}) = \frac{1}{2}(\boldsymbol{v}-\boldsymbol{v}^f)^T(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}-\boldsymbol{v}^f) + \frac{1}{2}(\boldsymbol{v}-\boldsymbol{y})^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{v}-\boldsymbol{y}). \tag{6.29}$$

The minimum to this equation yields our iterative update scheme. This allows us to construct a new prediction $\boldsymbol{v}$ given a first-guess $\boldsymbol{v}^f$ and an observation $\boldsymbol{y}$.

6.1.5 Deriving an iterative update

We then minimise the variational functional

$$\mathcal{J}[\boldsymbol{v}^a] = (\boldsymbol{v}^f-\boldsymbol{v}^a)^T(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}^f-\boldsymbol{v}^a) + (\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a) \tag{6.30}$$

so that we have a weighted contribution from the forecast of the state-space model and the observations. We recognise that minimising Equation 6.29 with respect to $\boldsymbol{v}$ gives the estimate $\boldsymbol{v}^a$ with assumed Gaussian priors. We then minimise the functional by setting the variation in terms linear in $\delta\boldsymbol{v}^a$ to zero. Doing so yields

$$\begin{aligned}
\delta\mathcal{J} &= \delta\mathcal{J}[\boldsymbol{v}^a+\delta\boldsymbol{v}^a] - \mathcal{J}[\delta\boldsymbol{v}^a],\\
&= -(\delta\boldsymbol{v}^a)^T(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}^f-\boldsymbol{v}^a) - (\boldsymbol{v}^f-\boldsymbol{v}^a)^T(\boldsymbol{C}^f)^{-1}\delta\boldsymbol{v}^a -\\
&\qquad (\boldsymbol{H}\delta\boldsymbol{v}^a)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a) - (\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a)^T\boldsymbol{\Gamma}^{-1}\boldsymbol{H}\delta\boldsymbol{v}^a + \mathcal{O}((\delta\boldsymbol{v}^a)^2).
\end{aligned}$$

We now minimise with respect to $\delta\boldsymbol{v}^a$. This gives

$$-(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}^f-\boldsymbol{v}^a) - (\boldsymbol{v}^f-\boldsymbol{v}^a)^T(\boldsymbol{C}^f)^{-1} - \boldsymbol{H}^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a) - (\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a)^T\boldsymbol{\Gamma}^{-1}\boldsymbol{H} = 0$$

We now note that the error covariance matrices $\boldsymbol{C}^f$ and $\boldsymbol{\Gamma}$ are symmetric and positive-definite. Given this, it is possible to show that $(\boldsymbol{C}^f)^{-1}$ and $\boldsymbol{\Gamma}^{-1}$ are also symmetric and positive definite. We can also deduce that $\boldsymbol{A}^T\boldsymbol{B} = \boldsymbol{B}\boldsymbol{A}$ if $\boldsymbol{B}$ is symmetric. So we get

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \boldsymbol{C}\boldsymbol{H}^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{v}^a) \tag{6.31}$$

We now define $\boldsymbol{r} = \boldsymbol{H}\boldsymbol{C}^f$ as the error covariance functions for measurements, and $\boldsymbol{b} = \boldsymbol{\Gamma}^{-1}\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^a$ to give

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \boldsymbol{r}^T\boldsymbol{b}. \tag{6.32}$$

Solving for the coefficients $\boldsymbol{b}$ using the definition of $\boldsymbol{b}$ and the above equation gives

$$\boldsymbol{b} = \boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{v}^f + \boldsymbol{r}^T\boldsymbol{b}))$$

which can be rearranged for the solvable equation for the coefficients of $\boldsymbol{b}$

$$(\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{H}^T + \boldsymbol{\Gamma})\boldsymbol{b} = \boldsymbol{d} - \boldsymbol{H}\boldsymbol{v}^f.$$

Explicitly, this results in

$$\boldsymbol{b} = (\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{H}^T + \boldsymbol{\Gamma})^{-1}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{v}^f)$$

for $\boldsymbol{b}$. Substituting this expression for $\boldsymbol{b}$ into Equation 6.32 gives

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \boldsymbol{C}^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{H}^T + \boldsymbol{\Gamma})^{-1}(\boldsymbol{d} - \boldsymbol{H}\boldsymbol{v}^f) \tag{6.33}$$

where it is more common to define the Kalman gain matrix $\boldsymbol{K}$ such that

$$\boldsymbol{K} = \boldsymbol{C}^f\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{H}^T + \boldsymbol{\Gamma})^{-1} \tag{6.34}$$

and the original correction equation becomes

$$\boldsymbol{v}^a = \boldsymbol{v}^f + \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^f). \tag{6.35}$$

We can also derive an update for the covariance matrix $\boldsymbol{C}^a$ by noting that

$$\begin{aligned}\boldsymbol{C}^a &= \mathbb{E}[(\boldsymbol{v} - \boldsymbol{v}^a) \otimes (\boldsymbol{v} - \boldsymbol{v}^a)],\\ &= \mathbb{E}[(\boldsymbol{v} - \boldsymbol{v}^f - \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^f)) \otimes (\boldsymbol{v} - \boldsymbol{v}^f - \boldsymbol{K}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^f))]\end{aligned}$$

Now using that $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{v} + \boldsymbol{\eta}$ we can gather like terms and write

$$\begin{aligned}\boldsymbol{C}^a &= \mathbb{E}[((I - \boldsymbol{K}\boldsymbol{H})(\boldsymbol{v} - \boldsymbol{v}^f) - \boldsymbol{K}\boldsymbol{\eta}) \otimes ((I - \boldsymbol{K}\boldsymbol{H})(\boldsymbol{v} - \boldsymbol{v}^f) - \boldsymbol{K}\boldsymbol{\eta})],\\ &= (I - \boldsymbol{K}\boldsymbol{H})\mathbb{E}[(\boldsymbol{v} - \boldsymbol{v}^f) \otimes (\boldsymbol{v} - \boldsymbol{v}^f)](I - \boldsymbol{K}\boldsymbol{H})^T + \boldsymbol{K}\mathbb{E}[\boldsymbol{\eta} \otimes \boldsymbol{\eta}]\boldsymbol{K}^T,\\ &= (I - \boldsymbol{K}\boldsymbol{H})\boldsymbol{C}^f(I - \boldsymbol{K}\boldsymbol{H})^T + \boldsymbol{K}\boldsymbol{\Gamma}\boldsymbol{K}^T \end{aligned} \tag{6.36}$$

Now, from the definition of $\boldsymbol{K}$, use that $\boldsymbol{C}^f\boldsymbol{H}^T = \boldsymbol{K}(\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{H}^T + \boldsymbol{\Gamma})$ in the above equation to get

$$\boldsymbol{C}^a = (I - \boldsymbol{K}\boldsymbol{H})\boldsymbol{C}^f = \boldsymbol{C}^f - \boldsymbol{r}^T(\boldsymbol{H}\boldsymbol{C}^f\boldsymbol{M}^T + \boldsymbol{\Gamma})^{-1}\boldsymbol{r} \tag{6.37}$$

and we obtain an update equation for the improved error covariance matrix in terms of the first guess error matrix. So we have derived a set of equations which allows for an improved prediction of $\boldsymbol{v}^a$ given $\boldsymbol{C}^f, \boldsymbol{\Gamma}, \boldsymbol{v}^f$ and $\boldsymbol{y}$. We then must be able to accurately estimate $\boldsymbol{C}^f$ at some time t_k. The sequential Kalman filter described above is suited for linear dynamics, given the assumption of Gaussian errors in its construction.

6.1.6 The Recursive Algorithm

Putting all the components together, we are left with the following iterative scheme:

1. Calculate the Kalman Gain $\boldsymbol{K}_j = \boldsymbol{C}_j^f \boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}_j^f \boldsymbol{H}^T + \boldsymbol{\Gamma})^{-1},$ (6.38)
2. Update State Estimate $\boldsymbol{v}_j^a = \boldsymbol{v}_j^f + \boldsymbol{K}_j(\boldsymbol{y}_j - \boldsymbol{H}\boldsymbol{v}_j^f),$ (6.39)
3. Update Covariance $\boldsymbol{C}_j^a = (\mathbb{I} - \boldsymbol{K}_j \boldsymbol{H})\boldsymbol{C}_j^f,$ (6.40)
4. Update to $j+1$ $\boldsymbol{v}_{j+1}^f = \boldsymbol{M}\boldsymbol{v}_j^a,$ (6.41)

$$\boldsymbol{C}_{j+1}^f = \boldsymbol{M}_j \boldsymbol{C}_j^a \boldsymbol{M}_j^T + \boldsymbol{Q}. \tag{6.42}$$

The iterative scheme for the ensemble Kalman filter will also work in a similar way. The update for the Kalman gain is sometimes instead written as the solution to the linear system of equations

$$\boldsymbol{K}_j \boldsymbol{S}_j = \boldsymbol{C}_j^f \boldsymbol{H}^T \tag{6.43}$$

where $\boldsymbol{S}_j = \boldsymbol{H}\boldsymbol{C}_j^f \boldsymbol{H}^T + \boldsymbol{\Gamma}$. This solution is preferred for numerical stability [107].

6.1.7 Nonlinear Extensions of the Kalman Filter

The Kalman filter is derived based on the assumption of a linear system of equations with associated normally distributed errors. In reality, many systems are nonlinear and the filter will perform poorly on these problems. This limitation has led to the development of many nonlinear extensions (see for example [155]). The first nonlinear extension developed is called the extended Kalman filter (EKF) which involves linearising the now nonlinear space-state model about the assimilation point and developing a tangent approximation to the nonlinear system. The Kalman update process is then evaluated with a Taylor expansion of the original nonlinear system. However, for most problems, there are better alternatives as this requires the linearisation to remain valid between assimilation time-steps and will inevitably fail for highly nonlinear systems. It also requires labour intensive evaluations of the Jacobian and analytic derivations of the expanded system of equations which are infeasible for larger systems of equations [69].

More recently, the unscented Kalman Filter (UKF) was introduced [47] to address nonlinear state-estimation. While in EKF, the covariance matrices are propagated through linearised forms of nonlinear update equations, in UKF covariance matrices are estimated by a careful selection of sigma points which are passed through the nonlinear update equations to estimate the updated covariance matrix. By carefully selecting these points, it is possible to accurately estimate the means and variance despite a nonlinear transformation being applied. These have not seen much application when the state-dimension is high as the number of sigma points scales with the dimension of the problem [69].

Here we primarily consider work following from previous students at CCFE (Luca Spinicci and Ana Osojnik) which revolved around using the ensemble Kalman filter for applications to nonlinear systems introduced by [24]. This involves estimations of the covariance matrices by calculating them as the average of an ensemble of members which are propagated by the nonlinear state operator. In some ways this allows us to ignore the Gaussian assumption in the prior [69], hence making it suitable for our applications. However, it's still worth remembering that the error statistics are represented by the first and second order moments and so we still use Gaussian assumption in the estimation of the posterior. Despite this, EnkF has seen much success in cases where these assumptions are violated [100]. The EnKF method bears a strong resemblance to the approach in UKF in which a specific selection of ensemble members are also propagated by the nonlinear space-state model. We will mention that EnKF is the simpler of the two methods to implement as we do not need to choose the location and weightings of the sigma points. However, this will typically mean that we require more ensemble members than UKF to accurately determine the covariance matrices.

6.1.8 The Ensemble Kalman Filter

While the linear Kalman filter is the provably best linear estimator, the update of the background covariance matrix $\boldsymbol{C}^f$ scales as $\mathcal{O}(n^3)$ making this approach unsuitable for large state-space dimension n. The ensemble Kalman filtering method (EnKF) overcomes this limitation by using a random-sampling Kalman filtering approach, where the covariance matrices are approximated from an ensemble of N state realisations. In this view, we no longer need to propagate the covariance matrices and instead take the mean of the ensembles as the best estimate and the spread of the ensemble members represents the error. Further, while the assumptions of Gaussian errors are still similar to the Kalman filter, practical success can be achieved applying EnKF to nonlinear state-space models and non-Gaussian noise, precisely because we no longer propagate the covariance matrix.

The method works by estimating the covariance matrices by evolving N ensemble members in time where typically $N < n$. Previously the covariance matrices were calculated as

$$\begin{aligned}\boldsymbol{C}_j^f &= \mathbb{E}[(\boldsymbol{v}^t(t_j) - \boldsymbol{v}^f(t_j)) \otimes (\boldsymbol{v}^t(t_j) - \boldsymbol{v}^f(t_j))],\\ \boldsymbol{C}_j^a &= \mathbb{E}[(\boldsymbol{v}^t(t_j) - \boldsymbol{v}^a(t_j)) \otimes (\boldsymbol{v}^t(t_j) - \boldsymbol{v}^a(t_j))].\end{aligned}$$

We now replace these with Monte-Carlo approximations

$$\boldsymbol{C}_j^f \approx \frac{1}{N-1} \sum_{i}^{n=N} (\boldsymbol{v}_i^f - \bar{\boldsymbol{v}}^f) \otimes (\boldsymbol{v}_i^f - \bar{\boldsymbol{v}}^f) \tag{6.44}$$

$$\boldsymbol{C}_j^a \approx \frac{1}{N-1} \sum_{i}^{n=N} (\boldsymbol{v}_i^a - \bar{\boldsymbol{v}}^a) \otimes (\boldsymbol{v}_i^a - \bar{\boldsymbol{v}}^a) \tag{6.45}$$

are the unbiased estimates of the correlation matrices and $\bar{\boldsymbol{v}}^f = \frac{1}{N}\sum_i^{n=N} \boldsymbol{v}_i^f$ is the mean which is now regarded as our best estimate. If each ensemble member were to see the same observation, the covariance would be underestimated as all trajectories would be corrected toward the same point. Following this, the procedure is similar to before, with the calculation of the Kalman gain matrix

$$\boldsymbol{K}_j = \boldsymbol{C}_j^f \boldsymbol{H}_j^T \left(\boldsymbol{H}_j \boldsymbol{C}_j^f \boldsymbol{H}_j^T + \boldsymbol{\Gamma}_j\right)^{-1} \tag{6.46}$$

followed by constructed a linear correction for each ensemble member

$$\boldsymbol{v}_i^a(t_j) = \boldsymbol{v}_i^f(t_j) + \boldsymbol{K}_j \left(\boldsymbol{y}(t_j) + \boldsymbol{e}_i(t_j) - \boldsymbol{H}_j \boldsymbol{v}_i^f(t_j)\right). \tag{6.47}$$

As discussed by [31], we now perturb the observations $\boldsymbol{y}_i$ by some value in $\boldsymbol{e}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Gamma})$ to account for known measurements in the ensemble framework. Finally, we can compute outputs based on the corrected mean

$$\bar{\boldsymbol{v}}^a(t_j) = \frac{1}{N} \sum_{i=1}^{n=N} \boldsymbol{v}^a(t_j) \tag{6.48}$$

and covariance analysis matrix

$$\boldsymbol{C}_j^a \approx \frac{1}{N-1}\sum_{i}^{n=N}(\boldsymbol{v}_a^i - \bar{\boldsymbol{v}}_a)\otimes(\boldsymbol{v}_a^i - \bar{\boldsymbol{v}}_a). \tag{6.49}$$

Like the unscented Kalman filter (see ref [47]), the EnKF method passes a number of ensemble members through the space-state model. The Monte-Carlo estimate for the covariance matrices thus often allows us to make accurate predictions of the mean and covariance of the system as we are using a sampling method, in contrast the regular Kalman filter. As the covariance matrices are evolved through the space-state model, even when the model is linearised, we often end up with completely incorrect estimates of the mean and covariance. Explicit calculation of the covariance matrices can be entirely avoided in Equation 6.46 by instead solving a least-squares problem with the ensemble discrepancies [107] if computational performance is important.

6.1.9 Parameter and state estimation

In cases where parameter estimation is of interest, we can simply modify v_j to include parameters $\lambda \in \mathbb{R}^L$ where $L = 1, \ldots, n_\lambda$ corresponding to a total of n_λ parameters. The state-vector can then be expressed as $\boldsymbol{x}_j = [\boldsymbol{v}_j^T, \boldsymbol{\lambda}_j^T]^T$. The observation vector on the other hand does not necessarily need to be modified, though can be if parameter measurements are available. However, given the increased dimension of $\boldsymbol{x}_j$, the observation operator must be modified to account for this. We then re-express $M_j = [\mathbb{I}_{n\times n}; \boldsymbol{0}_{n\times n_\lambda}]$ where ; represents vertical concatenation. As the Kalman filter progresses, we should improve the estimate of $\boldsymbol{\lambda}$. In the majority of parameter estimation problems it is common to assume that parameters remain approximately constant in time.

6.1.10 Optimisation of the initial conditions

One substantial issue with the current parameter and state estimation using the ensemble Kalman filter arises from sensitivity to the initial condition in the model. In general dynamical systems can exhibit starkly different behaviours even due to small changes in initial conditions or parameters. Previous work by earlier students [111] focused around obtaining better initial estimates for the state vector used in the Kalman filtering method.

To create a set of estimates for the initial conditions, the mean squared error between the model forecast $\boldsymbol{y}_j^f$ and the observations $\boldsymbol{y}_j$ between $[t_0, t_J]$ is minimised. In other words we aim to find $\boldsymbol{\lambda}$ and $\boldsymbol{x}_0$ such that

$$\underset{\boldsymbol{\lambda},\boldsymbol{v}_0}{\arg\min}\sum_{j=0}^{J}||\boldsymbol{y}_j^f - \boldsymbol{y}_j||^2 = \underset{\boldsymbol{\lambda},\boldsymbol{v}_0}{\arg\min}\sum_{j=0}^{J}||H^n(M_{\boldsymbol{\lambda}}^n(\boldsymbol{v}_0)) - \boldsymbol{y}_j||^2 \tag{6.50}$$

where n corresponds to the Runge-Kutta timestep. In reality, while we are estimating $\boldsymbol{v}_0$ and $\boldsymbol{\lambda}$, these are expressed as one single state vector or column vector $\boldsymbol{x}$. Thus the problem is simply a minimisation in $\boldsymbol{x}$. To summarise, we aim to construct an estimate of the parameters $\boldsymbol{\lambda}$ and $\boldsymbol{v_0}$ by minimising the forecast over J observations. We further define the sampling rate per period of oscillation to be given by ν. The total number of observations is then defined over only one period of oscillation by $J = \nu - 1$.

Finally, to ensure the system displays stable behaviour, constraints are imposed on the parameter σ in the ANAC4 system. For this set of ODEs, stable oscillatory behaviour occurs for all $\boldsymbol{x}_0$ when $\sigma < 0$. During the optimisation, an upper bound of $\hat{\sigma} = 10^{-16}$ is placed on this coefficient. Three different optimisers are compared in the solution to the minimisation problem:

1. a derivative-free Gauss-newton solver (DFOGN) [121]
2. DFOGN with approximate model interpolation and restarts (DFO-LS) [120]
3. Covariance Matrix Adaption Evolution Strategy (CMA-ES) [125]

Here we will not go into detail how each of the optimisers work, and instead refer the reader to the relevant papers. We note that while this was implemented in the implementation of EnKF described in this thesis, it was not used and validation results have already been given by ref [106].

This type of optimisation can be challenging, as by the nature of EnKF we expect the parameters to eventually converge to their true values and this does not need to necessarily happen with the first J observations. Further, the specification of the initial state uncertainty is inevitably important to the convergence of such an optimisation and without any prior knowledge of the parameters we also have no knowledge of the uncertainty. A large initial uncertainty will then likely not provide good convergence in the optimisation method. Finally, as will be seen in later examples with the ANAET model, parameter convergence may be on slow time-scales of the system which would result in quite an expensive optimisation process. However, it could still be beneficial to explore this approach in more detail as most of the time we have no real knowledge of expected parameters.

6.1.11 Covariance Inflation and Error Influence

Inflation methods are used to describe a set of approaches designed to mitigate against errors which are not modelled during filtering. These can come from a wide range of processes such as sampling error due to finite ensemble sizes, unrepresented errors in the chosen model or incorrect estimates of measurement noise. In general, sampling errors are important when the prior cannot sufficiently represent the entirety of the phase space ($N \ll n$) and even cases with $N > n$ [107]. Sampling errors like this lead to an underestimation in state uncertainty and spurious correlations in variables. In our case we generally try to ensure N is sufficiently large that sampling errors are largely avoided. Another source of error is known as model errors. For example, assimilation can cause an over-confidence in the space-state model for chaotic systems which effectively ignore future measurements and eventually diverge from the true solution and model error must be included to prevent this. In parameter estimation problems it is common to assume that the dominant errors result from uncertainty in the parameters [69].

Ensembling methods are generally remarked to underestimate uncertainty matrices due to their sampling based approach. Covariance inflation, first introduced by [32], gives an approach to mitigate against this for a small number of ensemble members. For each ensemble member we increase the ensemble spread by multiplying the ensemble discrepancy by a constant c

$$\tilde{\boldsymbol{v}}_b^i = c(\boldsymbol{v}_b^i - \bar{\boldsymbol{v}}_b). \tag{6.51}$$

which is also referred to as multiplicative inflation [155]. We would remark that this inflation is sometimes applied at different stages to counter different sources of error, and is commonly applied on assimilation time-steps. When applied on assimilation time-steps regions of dense measurements will cause more inflation of the ensemble members and may or may not be desirable [67]. Conversely it may be applied at integration time-steps of the model and thus viewed as a constant source of model error which does not inflate more if observations are dense. Adaptive schemes have also been proposed by [70] where the inflation factor is estimated jointly as a state variable. This requires some modification of covariance calculations to circumvent closure issues.

The point of application of inflation also alters which error it will mitigate against. If inflation is applied to the forecast ensemble, this can be viewed as a method of counteracting model error. If it's applied to the analysis ensemble then this is an attempt to mitigate errors in the analysis stage of EnKF [155]. When applied to the analysis members, different approaches exist such as the relaxation to prior perturbations (RTTP)

$$\boldsymbol{v}_{a,c}^{i} = (1-\alpha)\boldsymbol{v}_{a}^{i} + \alpha\boldsymbol{v}_{b}^{i} \tag{6.52}$$

written as a linear combination of the analysis and prior ensemble members for some constant $0 \leq \alpha \leq 1$ [50]. This has the effect of reducing the convergence of the ensemble members by relaxing the analysis members closer to the prior. The relaxation is also proportional to reduction in variance of the assimilation of an observation. This view of multiplicative inflation is actually a combination of both additive and multiplicative inflation but does have a reformulation in terms of a relaxation of the ensemble standard deviation to prior which is purely multiplicative [91]. A comparison of the two different approaches is given by [91], who suggest that additive inflation is more important when model error dominates and multiplicative inflation is more relevant when sampling errors dominate.

An alternative view of inflation is as an additive perturbation to the ensemble forecast which is also referred to as the process noise. The exact construction of this for the linear Kalman filter depends on the discrete form of the space-state model. For the ensemble Kalman filter, the appropriate choice of process noise $\boldsymbol{Q}$ is not particularly clear and there is no one-size fits all argument for its design. It typically depends on the source of error for the given problem at hand. A more comprehensive discussion of tuning inflation factors has been discussed by [38] and [51].

Both additive and multiplicative inflation have been used prior in the CCFE code with some reports of success (see [115]). The additive inflation in this code is referred to as a Gaussian kick to the ensemble members which is equivalent to process noise. The chosen multiplicative scheme applies a scaling to the forecast ensemble members and thus can be viewed as having a similar role as the additive inflation used by the CCFE code.

6.2 EnKF and symmetry-based models of gross tokamak behaviour

We now apply the outlined EnKF approach to create state and parameter estimates of the ANAC and ANAET models derived in § 2.4.2. These models are intended to match experimentally observed instabilities in a tokamak, in particular sawtooth oscillations. To work towards a future application to experimental data we first consider the challenges of modelling the simpler ANAC model which totals 4 unknown predictions; two for the state and two for the parameters. Observations will be generated from integration of the ANAC model itself with the data subsequently down-sampled and noise added to each observation. We then consider modelling the ANAET model with either 9 or 11 state variables which represents a significantly more challenging case.

When working towards experimental data, we note some of the following expected characteristics to aim for

1. The data contains noise from multiple sources including: additive noise, multiplicative noise and truncations in recorded data points. The degree of noise is assumed to generally be ±10% of the signal [17]. To simplify the analysis initially we only assume additive noise.
2. we assume that the sampling rate of the signal is sufficient to determine the period of the signal through a FFT and that there are at least 6 samples per period of the smallest scale in the measurement

3. presently we assume that the signals are stationary, though this is not necessarily true however the signals could be trend filtered to resolve this [89]
4. Only partial observations of system are available but those measurements are direct implying $\boldsymbol{H} = \mathbb{I}$ in the EnKF method.

In the present work we primarily limit ourselves to sparsely sampled data with varying degrees of observational noise.

EnKF represents a useful tool in dealing with these outlined challenges, in particular because of how it treats noise. Firstly, EnKF is a derivative free approach as the updated state estimation comes from the minimisation of Equation 6.30 through the calculation of the Kalman gain. This provides an excellent robustness to noise. Second, specifying the process noise allows the modelling of both stochastic processes but to accommodate for unrepresented model errors (provided they are Gaussian). While the assumptions of Gaussian errors may seem stringent, in practice this can, and will often be violated in EnKF and the method will still perform well. As discussed this is in part due to the Monte-Carlo estimation of the covariance matrices.

Further, EnKF also allows us to model unobserved variables as we assumed in the construction of our system that the observations could be fewer than the number of modelled variables. While our system will consist of direct observations of the variables, EnKF allows for combinations or transforms of the observations to be applied. Finally the ensembling approach allows us to consider nonlinear systems.

6.2.1 Terminology

There are a number of terms which are used throughout use of the code which are documented here

1. dt the integration time-step of the space-state model
2. dt_{assim} The assimilation time-step which is usually larger than dt
3. δ_{obs} the standard deviation of the additive Gaussian noise

6.2.2 Example with previous code

dt	dt_{assim}	μ	σ	a_0	$\dot{a}_0$	δ_{obs}	observed
0.005	0.005*48	10	-0.1	1.5	0	0.1	$a,\ \dot{a}$

Table 3: Table of parameters for the observations.

We reproduce some results of the previous code adapted by [115] to ensure the function of the pre-existing software. All examples in this reference make use of the ANAC model

$$\ddot{a} = \mu a + \sigma a^3 \tag{6.53}$$

which is sometimes referred to as the ANAC4 model when both parameters and a states are being estimated (4 total estimates required of EnKF). As the code is several years old, some functionality had to be updated. We begin by running example scripts left which reproduce results with the ANAC model. To run the example, we first begin by generating observations using `generate_data.py`. In this instance, we assume that the assimilation occurs at an integer time-step of the integration. To generate this data, the model is first integrated in a deterministic fashion using the true initial conditions $\boldsymbol{x} = (a, \dot{a}, \mu, \sigma) = (1.5, 0, 10, -0.1)$ at a time-step of $\mathrm{d}t = 0.005$. This time-step is chosen to ensure sufficient accuracy with the Runge-Kutta solver. Note that the noise is not added at this stage of the example.

Following this, the initial conditions must be set either by directly specifying from, for example, the observation parameters in Table 3. Alternatively, the optimisation process described previously may be used. For this instance we use previously optimised initial conditions and plot a representative set of results in the following section.

6.2.3 EnKF within external Python packages

While the code used in the work by [115] works, there are some limitations listed below.

1. The code was originally adapted from Matlab code meaning the EnKF class is only used to saved parameters which are then imported repeatedly at different stages as is often the case in Matlab codes. This is not the intended use of Python classes.
2. Multiple imports of parameters defined in different places within the code, where it is not clear that parameters are actually used in certain aspects of the code function.
3. Boolean switches to define methods used are hard to keep track of.
4. Space-state models need to be manually included in separate scripts and parameter files written.
5. The measurement matrix cannot be easily defined within the parameters script.
6. Read and write function calls within the filter code which save in unclear locations and during assimilation slowing the function of the code.
7. Two different use cases of the code depending on artificial data and experimental data which can be consolidated to one base code.
8. In general it has a complicated code structure which make debugging and profiling challenging as input parameters are read from several different scripts.

Given the above points, we adopt the external python package `filterpy` and recreate the results and function of the CCFE EnKF code. This package has multiple Kalman filter flavours which can all be used within the same shared function calls as EnKF. The package also features an excellent accompanying book available interactively on Github [137] which makes use of `filterpy` to understand Kalman and Bayesian filters from an intuitive standpoint. It should be noted though that any commits have ended for this textbook and the Ensemble Kalman Filter is mentioned only in the Appendix as an additional feature. While the code has not been extensively validated, the assimilation procedure is identical to the process used in the pre-existing CCFE software. A fully worked example of creating an EnKF method is also available at the GitHub repository [141]. Benchmarking of different data assimilation methods is available at the repository [140] with corresponding notes [161]. For applications of the code discussed in this thesis, examples are given in the Github repository which allow the results to be reproduced `https://github.com/royalasdaircoding/enkf_ANAC.git`.

The class structure of `filterpy` allows us to address all of the issues listed above, making the code more readable. Further, optimisation is still possible using the publicly available DFO-GN [121] and DFO-LS [120] Python packages from the same authors as well as many other options available in packages like `NLopt`. We have left a number of scripts using a modified version of this package which leave examples on how to reproduce and extend prior work by CCFE students.

There are some limitations to the use of `filterpy`. While in theory parallelisation over ensemble members should be straightforward and an obvious source of speed up, this is rarely simple in Python. This is particularly true for a small number of ensemble members. The reason is that initiating CPU workers (depending on the parallelisation software) can involve copying the instance of the code several times and results in a large computational expense to begin with. Given the current structure of the code, the parallelisation would have to occur only one time, with communication between cores whenever covariance matrices are calculated. However, given the small dimension of the problem it's not clear that there are really significant gains to be made. Really, the best speed ups would come from not using Python entirely or at least making use of some strict typing through `cython`. Several issues occur when using Python. One example which will be relevant later is the use of the `scipy` [146] package `solve_ivp` to solve differential equations. When creating this function, this can involve invoking the Python interpreter which is a slow process. As each ensemble member must be integrated over assimilation time-steps, constantly invoking the Python interpreter results in a significant impact on performance. This will be discussed again when looking at stiffer ODEs.

6.2.4 Extensions to EnKF

Feature	Code Call	Purpose
Gauss Kick (GK)	`Gauss_kick`	Uniform Gaussian noise correcting for model error
Pre inflation	`inf`	Inflates the forecast ensemble members
Analysis inflation	`inf_a`	Inflates the analysis ensemble members
Constraints	`constraints`	Lower and upper bounds to be constrained
Prior type	`ensemble_type`	Uniform or Gaussian priors
Generator random numbers	`seed`	Allows local scoping of random seeds in EnKF class

Table 4: Additional EnKF class attributes that have been added to `filterpy`

While using the `filterpy` package, a number of changes were made and listed in Table 4 with their corresponding function calls in the EnKF class. Some of the features added are previous features included in the CCFE code such as Gauss kick, pre-inflation and prior type. Of the remaining features, analysis inflation and constraints are both newly added. Finally the generator random numbers exist so a seed can be set for reproducibility.

6.2.5 Notational Differences between `filterpy` and thesis

As we adapt an external package there are some notational differences. These are as follows:

1. $\boldsymbol{P}$ is a running calculation of the state error covariance matrix and can refer to either $\boldsymbol{C}_j^f$ or $\boldsymbol{C}_j^a$. When we discuss the calculation of the confidence intervals, when an observation is assimilated we calculate $\boldsymbol{C}_j^a$, otherwise we calculate $\boldsymbol{C}_j^f$.
2. $\boldsymbol{R}$ is the same as $\boldsymbol{\Gamma}$ representing the measurement noise covariance matrix.
3. $\boldsymbol{x}$ refers to the state variables

6.2.6 Results from the new Kalman software

We first test the new EnKF code on the ANAC4 model at the same parameters as the previous EnKF software. The structure of the code is simple and readable, we will include an example of the function of this code. We start with a set of import statements shown in Listing 2.

```python
from filterpy_local.kalman import EnsembleKalmanFilter as EnKF
from filterpy_local.common import Q_discrete_white_noise

import numpy as np
import matplotlib.pyplot as plt
import matplotlib
from scipy.integrate import solve_ivp
from numba import jit
```

Listing 2: Import statements for the Kalman filter code.

To create a space-state model of the EnKF code, we would also like to generate estimates for the parameters μ and σ. We can essentially include these variables which we expect to be constant in our dynamical system as follows

$$\begin{aligned}
\dot{a} &= v, \\
\dot{v} &= \mu a + 2\sigma a^3, \\
\dot{\mu} &= 0, \\
\dot{\sigma} &= 0.
\end{aligned}$$

Here we have assumed the simplest possible form for the parameters, that they are constant in time with no added noise. To generate a set of observational data we set the parameters to $\mu = 10$ and $\sigma = -0.1$ with variables set to $a_0 = 1.5$ and $v_0 = 0.5$. In this example, we also use the ANAC model as the space-state model for the EnKF routine. This represents an ideal scenario in which the only errors are those from the observations.

Further, we take the observations to be generated at a sampling rate of approximately $\nu = 12$ and the generated observations are downsampled accordingly. The period of the system is based on estimates using an FFT included in the previous CCFE code. The code for this process is shown in Listing 3. Creating the observational data is now a simple process whereby an integration method is specified, and noise is added post integration to the observations. Previously generating the clean observational data is done beforehand, and then noise is added by a separate function call when the main code is run. There are also different calls for experimental data and toy data. This approach avoids the need to have separate code function in the main EnKF software as observational data can be read from a separate file before instantiating the class.

```
#Define the ANAC model
@jit
def anac(x, dt):
    return np.array([x[1], x[2]*x[0] + 2*x[3]*x[0]**3, 0, 0])

@jit
def runge_fx(x, dt):
    k1 = anac(x, dt)
    k2 = anac(x + 0.5*dt*k1, dt)
    k3 = anac(x + 0.5*dt*k2, dt )
    k4 = anac(x + dt*k3, dt)
    x = x + (dt/6.0) * (k1 + 2*k2 + 2*k3 + k4)
    return x

# Generate synthetic data
def generate_data(steps, x0 = np.array([1.5, 1.,10., -0.1]), dt_phase_model=0.005):
    np.random.seed(1234)
    observations = np.zeros((steps,x0.shape[0]))
    observation_times = []
    tcounter = 0
    for i in range(steps):
        x0 = runge_fx(x0, dt_phase_model)
        tcounter+=dt_phase_model
        observation_times.append(tcounter)
        observations[i] = x0 + np.random.normal(0,std, size=x0.shape[0])
    return observations, np.array(observation_times)

true_params = [10, -0.1]
std = 0.1 #standard deviation of additive noise
# Parameters
dt_phase_model = 0.005
steps = 10000

# Generate synthetic data
observations, observation_times = generate_data(steps, dt_phase_model=
    dt_phase_model)

plt.figure(figsize=(12,8))
plt.plot(observations[:,0], observations[:,1], "b")
plt.xlabel(r"$a$")
plt.ylabel(r"$\dot{a}$")

########### downsample the data
observation_sub = 32
period=2.07
observations= np.reshape(observations[::observation_sub,1],
                         (observations[::observation_sub,1].shape[0], 1))
observation_times = observation_times[::observation_sub]
observational_nu =int(period//(dt_phase_model*observation_sub))

print("Observational data characteristics")
print("Period", period)
print("Samples per period", observational_nu)
```

Listing 3: Generate the observational data with added noise. Note this previous section of code spanned at least 4 separate scripts.

We next instantiate the EnKF class with parameters shown in Listing 4. We also perturb the true initial conditions with randomly generated noise from a uniform distribution between $[-2, 2]$ on each state variable mean. The observation function measurement $h(x)$ returns measurements of

a and v. The main function of the EnKF routine is then defined within the for loop where we choose to assimilate observations if they are close to the time-step of the integration regime. At this step, the EnKF routine makes a prediction of the ensemble mean and covariance matrices, correcting the resulting prediction depending on the observed measurement of $\dot{a}$. We also note that the observational data is sliced so that the observation at $t = 0$ is not used. There is also an ability to specify the process noise matrix. For now this is set to zero, but will be discussed in more detail later.

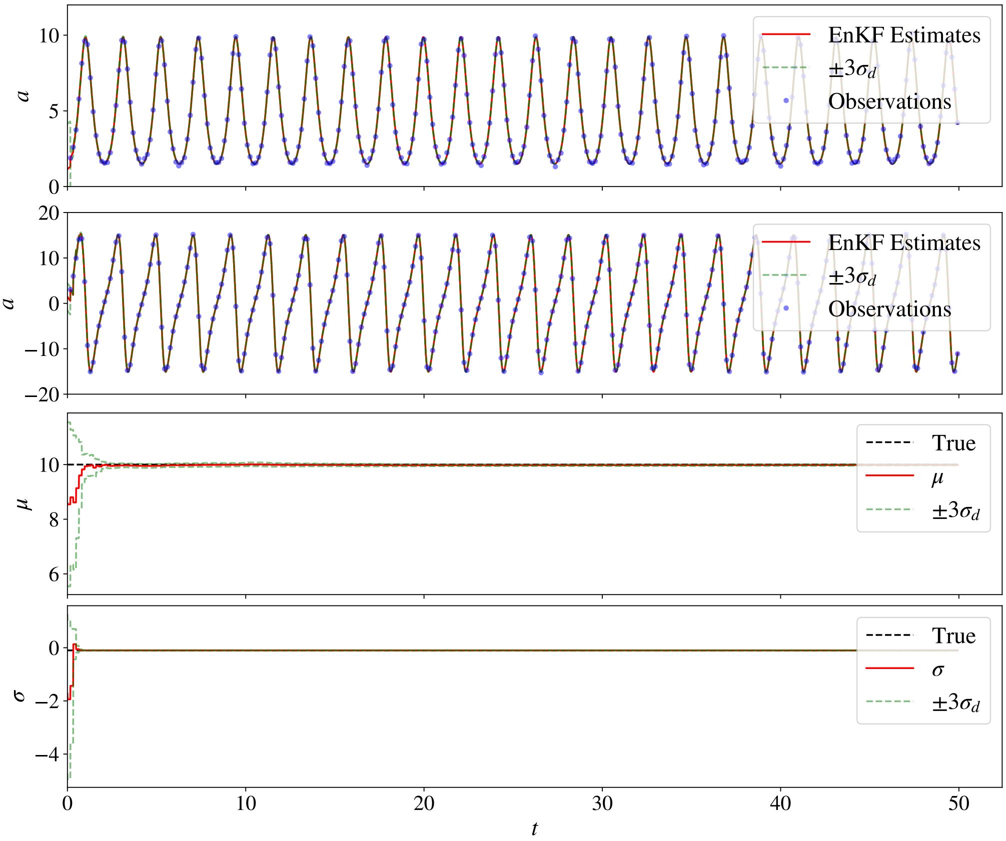


Figure 93: Convergence of the state and parameters for measurements of a and $\dot{a}$ using `filterpy`. The dashed green lines show the confidence intervals.

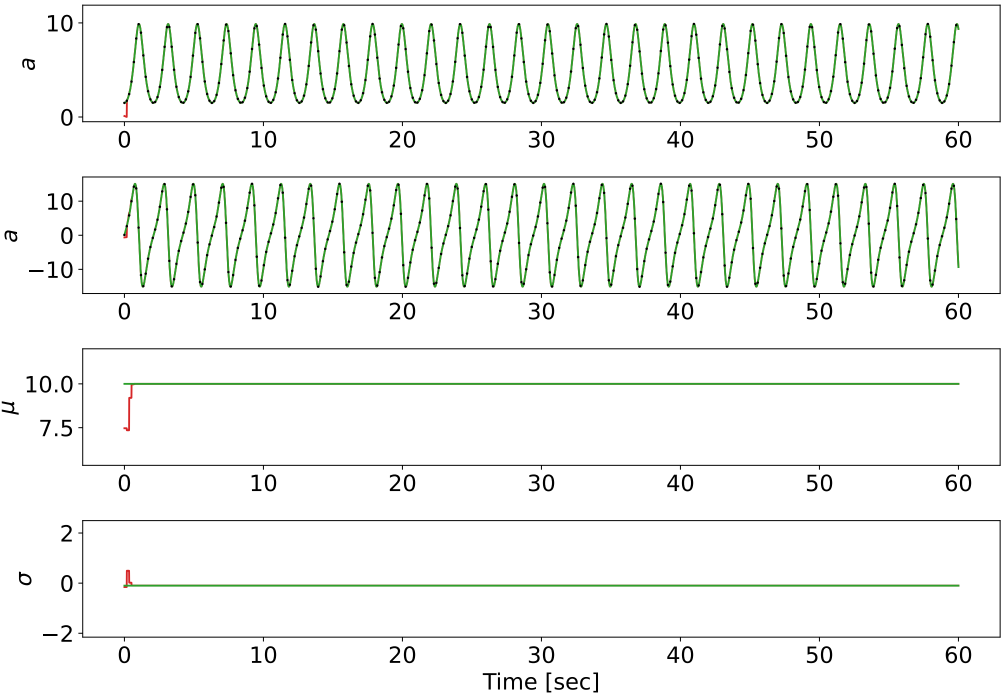


Figure 94: Example of the CCFE ENKF code assimilating using the same parameters over a similar observational data-set. Note the initial conditions used in this code are different and come from optimisation of the software. The green lines here represent the true data, black dots the observations and red are the ensemble averages. The dashed green lines show the confidence intervals.

```
#define the observation operator (only adot)
def hx(x):
    return [x[0], x[1]]

np.random.seed(1)
initial_covariance = np.eye(4) * 1
 initial_covariance[-1, -1] = 1
perturb = np.random.uniform(low=-2, high=2, size= 4)

initial_state = np.array([1.5, 0.5, 10, -0.1]) + perturb
 # perturb initial state and parameters
print("True initial condition")
print([1.5, 1,10., -0.1])
print("Perturbed initial condition")
print(initial_state)

# Ensemble Kalman Filter setup
enkf = EnKF(x=initial_state, #initial mean
            P=initial_covariance, #initial uncertainty
            dim_z=2, #number of observations
            dt=dt_phase_model, #integration time-step (not really used)
            hx=hx, #observation function
            N=30, #number of ensemble members
            fx=runge_fx, #nonlinear SSM
            ensemble_type="gsn", #initial prior distribution
            seed = 1, #seed for enkf, you must also set the seed when generating the observations
           )

Q = np.zeros((4,4)) #set the process noise to zero as in the CCFE code
```

```python
enkf.Q = Q
enkf.R = np.diag([std**2, std**2]) #measuremement uncertainty

#data stores
time_prediction = np.arange(0,observation_times[-1]+dt_phase_model, dt_phase_model)
estimates = np.zeros((time_prediction.shape[0], initial_state.shape[0]))
estimates[0,:] = initial_state
Pmat = np.zeros((time_prediction.shape[0], initial_state.shape[0], initial_state.shape[0]))
Pmat[0,:,:] = initial_covariance
update_counter = 0

#perform filtering
for ii, t in enumerate(time_prediction[1:]) :

    #assimilate observations if they are considered contemporary
    if np.isclose(t, observation_times[1:], rtol =1e-8, atol=dt_phase_model*1e-8).any():
        enkf.predict(True)
        enkf.update(observations[np.isclose(t, observation_times, rtol =1e-8,
                                            atol=dt_phase_model*1e-8),:][0])
        update_counter+=1

    #else integrate the ensemble members forward in time
    else:
        enkf.predict(False)
    estimates[ii+1,:] = np.copy(enkf.x)  # Extract only state variables from the state vector
    Pmat[ii+1,:,:] = np.copy(enkf.P)
print("total time", time.time()-start_time)
print("updated", update_counter, "times")
print("Final covariance P matrix")
print(enkf.P)
```

Listing 4: Perform ensemble Kalman filtering with $N = 30$, $P = \mathbb{I}$, $Q = \mathbf{0}$, $R = 0.1 * \mathbb{I}$.

The output of the assimilation process is shown in Figure 93. We see that the EnKF method is able to quickly converge from a relatively poor guess to the true mean and model parameters. We also note that the covariance matrix for these parameters quickly converges signified by the narrowing confidence intervals. The bounds here are determined from the analysis covariance matrix representing the uncertainty after combining the prior and likelihood. For similar parameter sets, this compares favourably to the CCFE code shown in Figure 94.

This whole process repeats in bulk the code contained within previous work by CCFE, neglecting the inclusion of the Gaussian kick as this partly functions like process noise. The advantage here is that the full class structure of the EnKF routine allows a new space state model or observations to easily be added and the whole code to generate this requires less interpretation and will only save diagnostics the user chooses to save. The major downside currently is the need to recursively call the EnKF class which in its current state would prevent easily parallelising over different ensemble members as an obvious point of increasing the speed of the code. As we would want to initiate parallel cores as few times as possible, it would be best to integrate in parallel over the entire assimilation time-step and not the phase space model time-step. Despite this, the code runs within $\sim 1s$.

6.2.7 Spread of the initial ensemble uncertainty

When using EnKF, there are several initial assumptions we must make. In particular these are the noise apparent in our measurements represented by $\boldsymbol{R}$, the initial uncertainty $\boldsymbol{P}_0$ from which our ensemble members are drawn, the initial condition from which to start the EnKF method and the degree of process noise $\boldsymbol{Q}$ in the system. The performance of EnKF will be dictated by an appropriate choice of these parameters and we must always consider if we have suggested sensible choices. In many cases our experimental measurements will have some known errors associated to them and this can be used to estimate $\boldsymbol{R}$ to a fair degree of accuracy. In this section we will discuss the impact of choosing these terms in relation to the ANAC model.

The example in the previous section shows the assimilation of two-state measurements for the ANAC model in an ideal situation. While we perturb the initial conditions for a and v, our choice of measurement uncertainty $\boldsymbol{R}$ and initial uncertainty $\boldsymbol{P}_0$ reflect somewhat ideal choices. In particular when choosing $\boldsymbol{P}_0$ it is important that the initial uncertainty in our 4 variables gives a sensible estimate. If this is not the case, we will likely have poor convergence in our parameters. If $\boldsymbol{R}$ is incorrect, we are either overconfident or underconfident in our measurements and this can also result in poor estimates. Incorrect estimates for $\boldsymbol{R}$ simply mean that the state-space model can be slow to respond to measurements if $\boldsymbol{R}$ is too large, and too responsive to measurements if it is too small over-fitting to the noise. The selection of $\boldsymbol{R}$ will be discussed later in the single-measurement case.

If we consider the collection of ensemble members as a matrix denoted $\boldsymbol{A}^f$, then details given in [55][pg 163] list the following conditions on the collection of ensemble members

1. the ensemble realisations should be realistic and physical acceptable.
2. $\text{rank}(\boldsymbol{A}^f) = \min(n, N)$ meaning that ensemble spans an N dimensional space.
3. The ratio of the smallest to largest eigenvalues of this matrix (the condition number) should be small. This relates to the linear independence of the different ensemble members.

This highlights that the correct specification of the initial uncertainty cannot be ignored. Future consideration of the initial ensemble could implement sampling methods given by ref. [46] which generate ensembles based off the singular value decomposition of a large ensemble.

A particular issue in selecting $\boldsymbol{P}_0$ with the ANAC model is that we may select initial conditions and parameters which are non-physical due to assumed Gaussian priors. In particular, if σ is positive and large we can get rapid divergence of the filter as this no longer acts as a damping term. Work by [106] remarked that there is divergence of the filter when the parameters are perturbed by values from a uniform distribution in $[-2, 2]$. This can largely be attributed to a poor guess in σ which results in diverging solutions of the space-state model over assimilation time-scales.

There is also mention by [106] that there can be poor convergence of μ depending on the initial guess for the parameters. The convergence of μ strongly depends on the initial uncertainty in the ensemble and there is no direct mention of this being selected. For instance, if the perturbation comes from a uniform distribution $[-2, 2]$, we would like for the initial ensemble to reflect this uncertainty. While this is a uniform distribution, we could select that 98% of members are within 3 standard deviations $3\sigma = 2$ and so the variance is $\eta = \frac{2}{3}^2$. Requiring that the initial guess is this accurate is likely too strict for the method, as any value in $x_0 + \tilde{x}_0$ is equally likely for the uniform perturbation, but the tails of our initial ensemble are much less likely if we use a Gaussian ensemble initially. In general, if poor convergence of the parameters is seen, increasing this initial uncertainty is important. This behaviour is alleviated by using an ensemble from a uniform distribution such that an equal weighting is given to members in the initial ensemble.

This is illustrated in Figure 95, where $\boldsymbol{P}_0 = 0.5 * \mathbb{I}$ has been used an initial state uncertainty with initial condition $\boldsymbol{x}_0 = [1.5, 1, 8., 1.5]$. We can see that while we generate a promising prediction of the measured outputs, the convergence of μ is not as good. We can see simply from the $\pm 3\tilde{\sigma}_d$ intervals that the true model parameters sit on the edge of the confidence intervals, implying that it is quite unlikely for the true parameters to lie in this region. We also see that the initial guess for σ is positive, which generates poor estimates for $\dot{a}$ which are unstable. After assimilation, σ is corrected to be closer to the true value. Given the rapid convergence of the measured variables, this is a case which may be improved using covariance inflation. This is just one representative run, and depends on the seed used and the perturbed observations in the EnKF method and the mean of the initialised ensemble.

If we increase the uncertainty to $\boldsymbol{P} = 2 * \mathbb{I}$ as in Figure 96, we can see significantly improved convergence to the true parameters. While there is certainly use in optimising for initial conditions as certain selections of parameters can cause divergence of the filter, it is also important to consider the confidence intervals in the parameters. In our case expanding the uncertainty causes the ensemble members to take larger positive values of σ.

We can create a more challenging situation if we consider a random uniform perturbation to the initial condition between $[-10, 10]$ for $a, \dot{a}, \mu$ and $[-2, 2]$ for σ with $\boldsymbol{P} = \text{diag}(9, 9, 9, 1)$ we can still achieve promising convergence of the parameters as shown in Figure 97. We restrict the size of the perturbation on σ as large values of the cubic term typically cause divergence of the space-state model before EnKF can correct the initial guesses.

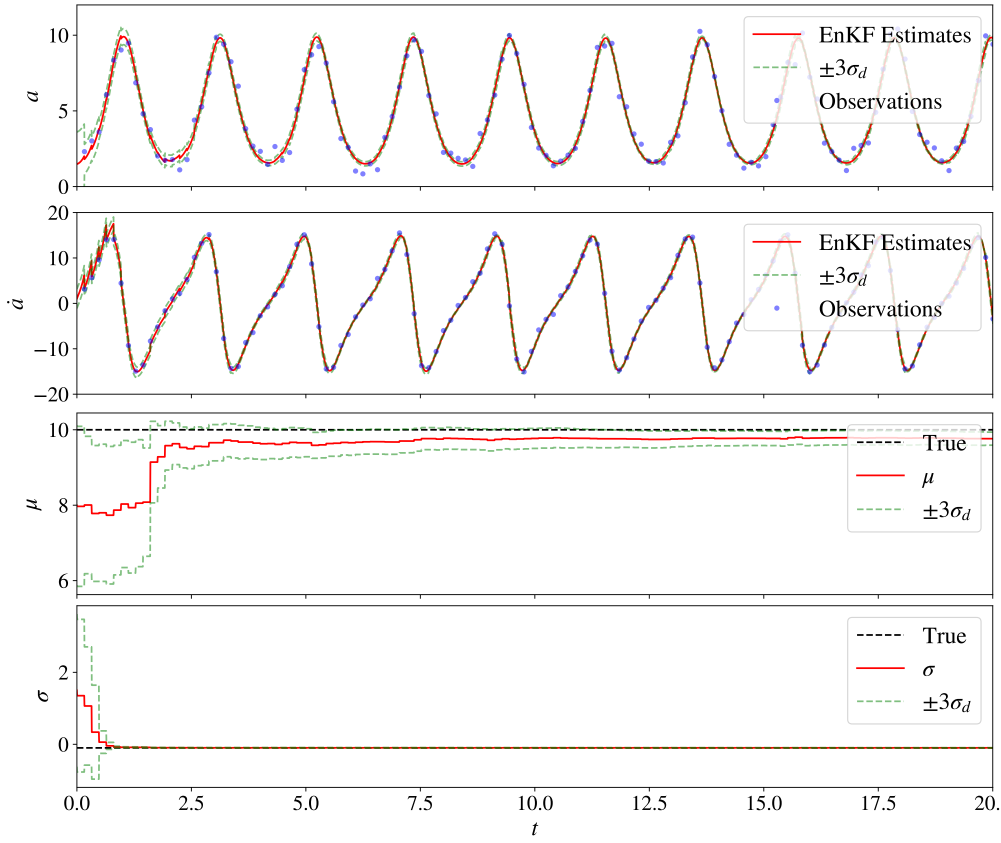


Figure 95: Convergence of the parameters with $\boldsymbol{P}_0 = 0.5 * \mathbb{I}$ and $\delta_{\text{obs}} = 0.5$. The dashed green lines show the confidence intervals.

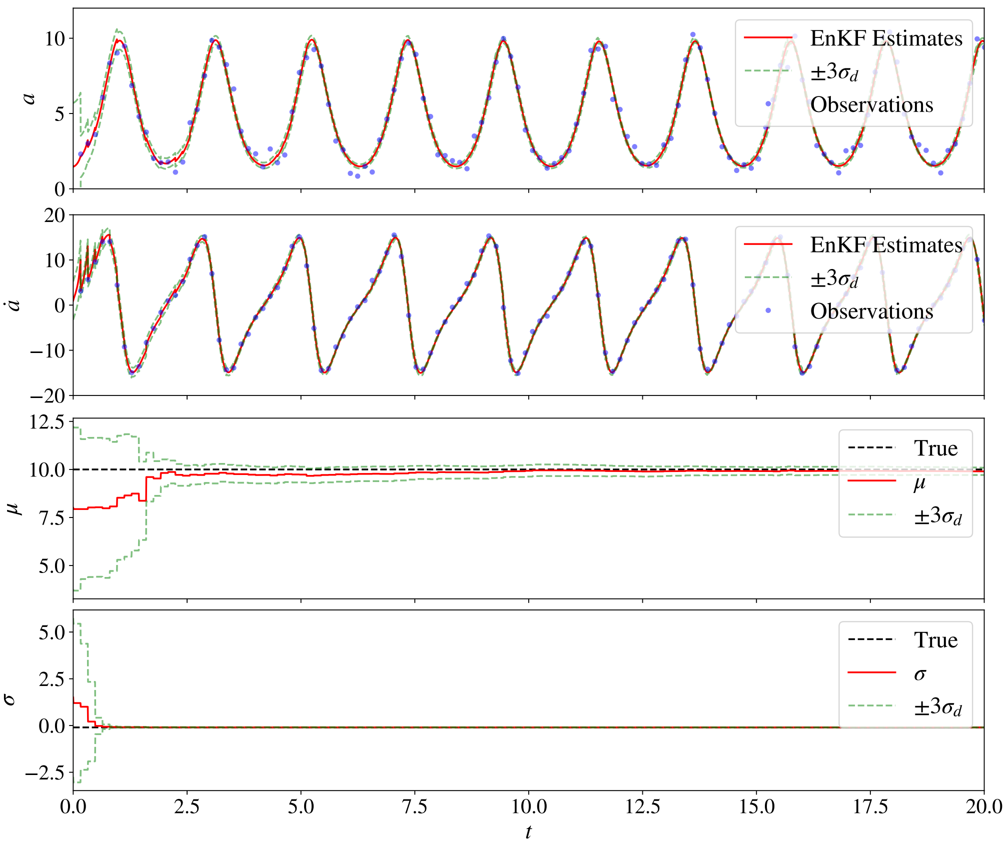


Figure 96: Convergence of the state variables and parameters with a large initial uncertainty $\boldsymbol{P}_0 = 2 * \mathbb{I}$ and observational noise with standard deviation $\delta_{\text{obs}} = 0.5$. The dashed green lines show the confidence intervals.

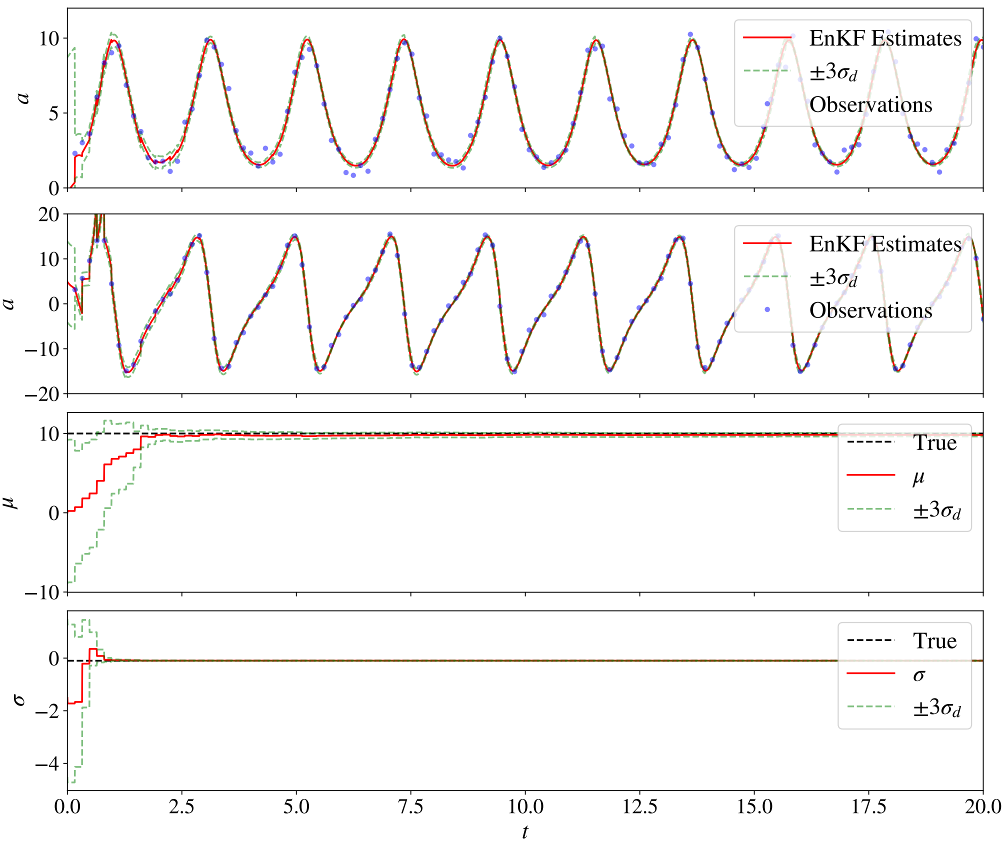


Figure 97: Convergence of the parameters with $\boldsymbol{P}_0 = \text{diag}(9, 9, 9, 1)$ and $\delta_{\text{obs}} = 0.5$ demonstrating good parameter convergence for very incorrect initial guesses. The dashed green lines show the confidence intervals.

6.2.8 Single measurement

An important extension toward application on experimental data is using EnKF in single measurement cases. The design of $\dot{a}$ is primarily intended to fit to Mirnov oscillations corresponding to MHD activity associated to neo-classical tearing modes or edge-localised modes [17, 84]. A genuine fit to experiment would only consider a single measurement of the rate of change of the poloidal field which we use to fit a surrogate model, the comparisons of which are made clearer in § 2.4.3. Here we will consider some challenges posed by fitting from measurements of $\dot{a}$ alone.

Previous work by [115] and [106] largely remained in the full state measurement case, though with some fitting to experimental data by [115] in a single-measurement to some filtered Mirnov signals relating to ELMs highlighting the viability of the method. In the latter case, a section of an experimental signal is considered and a simple harmonic oscillator is effectively identified with $\mu < 0$. Here we consider how the fitting performs generally in the single measurement case and whether the correlation in a and $\dot{a}$ is sufficient to infer the hidden states. We also consider how accurate we must be in our estimation of $\boldsymbol{R}$ to achieve this convergence, and discuss an alternative approach in parameter fitting compared to optimisation of initial conditions.

The code in Listing 4 can easily be modified to consider single measurements of $\dot{a}$ only. We again consider a case where we perturb the true initial condition x_0 by uniform random distribution between $[-2, 2]$, setting $\boldsymbol{P} = 2 * \mathbb{I}$. The results are shown in Figure 98 showing good convergence of the parameters in the single measurement case. Here we have reduced the number of periods over which prediction results are plotted for clarity.

We can note that $\boldsymbol{P}$ does converge quite rapidly so this problem may benefit from covariance inflation. However, it should not be surprising that we see convergence of covariance matrix as we assume no process noise and so are suggesting our space-state model should be an excellent predictor. The convergence of the confidence interval for μ is much slower than the other parameters, and this is a result of a small correlation between μ and $\dot{a}$. In general the convergence of the parameters is not as robust as the two state variable case.

In the single measurement case, we also tend to find that higher initial uncertainties can cause divergence of the filter and incorrect model parameters. We find that this can be resolved by using a higher number of ensemble members, suggesting that larger initial ensemble spread needs more ensemble members to accurately capture the mean and correct covariance. However, we emphasise that these improvements are not robust in the single measurement case and the filter can still often diverge. With larger initial uncertainties, increasing the number of ensemble members results in a higher chance of generating an ensemble member with poor initial guess which causes divergence. That being said, we still achieve promising convergence when the filter does not diverge.

Finally we note that convergence of the parameters takes longer than in the two measurement as in Figure 97. In the two measurement case, even with large initial uncertainties with high measurement noise, we can still get good convergence of the parameters within one period of oscillation. In the single measurement case, the convergence is not only harder, but takes longer and while the example shown occurs is around one period, often the filter diverges. We can see that this likely depends on our choice of initial uncertainty in σ shown in Figure 98, as some ensemble members will be initialised with positive values of σ. Overestimation of the state $\dot{a}$ can be seen at some points while the spread of these ensemble members lies within positive values of σ.

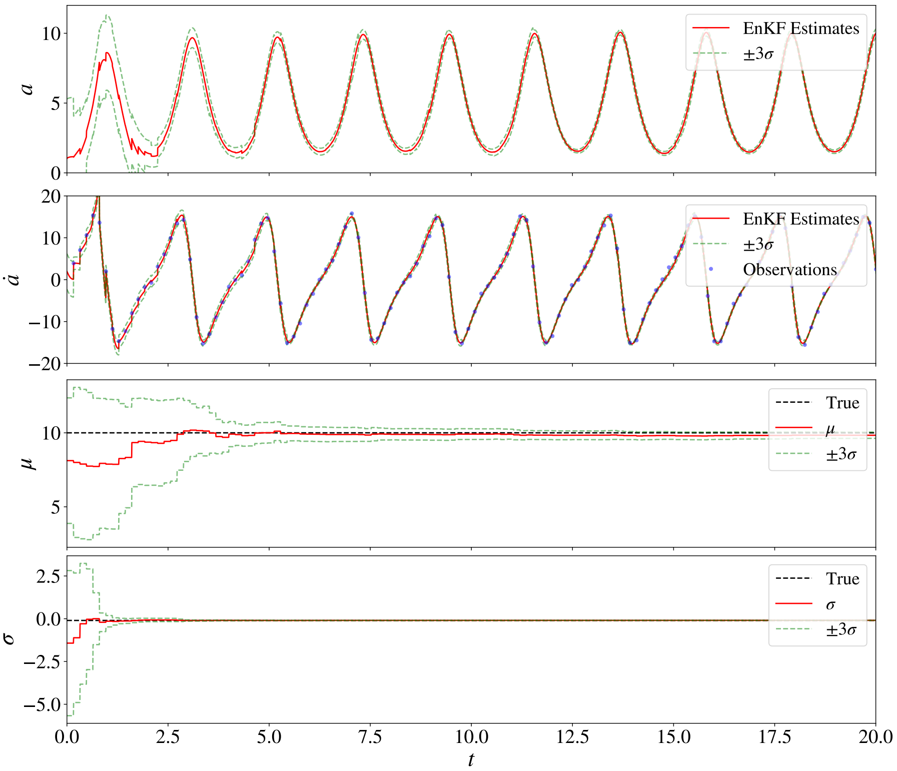


Figure 98: Convergence of EnKF with poor initial guesses with observational noise of standard deviation $\delta_{\mathrm{obs}} = 0.5$ and sampling rate $\nu = 12$ in the single measurement case. In this case the uncertainty bounds are reflective of the poor initial guesses.

6.2.9 Consideration of the explicit structure of the Kalman update with single measurements

Here we discuss briefly why we do not expect single measurement cases to perform as well for parameter convergence. Recall the Kalman gain is defined

$$\boldsymbol{K}_k = \boldsymbol{M}_k \boldsymbol{S}_k^{-1} \tag{6.54}$$

where the cross-covariance of the state and measurement matrix is given by

$$\boldsymbol{M}_k = \boldsymbol{P}_k \boldsymbol{H}_k^T. \tag{6.55}$$

If we consider explicitly the resulting form of this matrix for single measurements $\boldsymbol{H}^T = [0, H_2, 0, 0]^T$ then the result is

$$\boldsymbol{P}_k \boldsymbol{H}_k^T = \begin{bmatrix} P_{11} & P_{12} & P_{13} & P_{14} \\ P_{21} & P_{22} & P_{23} & P_{24} \\ P_{31} & P_{32} & P_{33} & P_{34} \\ P_{41} & P_{42} & P_{43} & P_{44} \end{bmatrix} \begin{bmatrix} 0 \\ H_2 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} H_2 P_{12} \\ H_2 P_{22} \\ H_2 P_{32} \\ H_2 P_{42} \end{bmatrix}. \tag{6.56}$$

The Kalman gain can then be written

$$\boldsymbol{K}_k = \begin{bmatrix} H_2 P_{12} \\ H_2 P_{22} \\ H_2 P_{32} \\ H_2 P_{42} \end{bmatrix} [H_2 P_{22} H_2 + R]^{-1} \tag{6.57}$$

Should any of these correlations between $\dot{a}$ and other variables be zero, there will be no corrections to the associated variable based on the observed measurements. Interestingly, this suggests all we care about in the single measurement case are correlations within these entries. This is of particular issue in the convergence of constant parameters μ and σ when incorrect initial guesses are used.

6.2.10 Incorrect selection of Measurement Uncertainty

So far $\boldsymbol{R}$ has been chosen based exactly on δ^2_{obs}. While this can be estimated beforehand, we would still like to consider the cases when this is estimated incorrectly. In the case where $\boldsymbol{R}$ is too small it will result in overconfidence in the measurements and the ensemble members will converge too quickly possibly resulting in incorrect estimates of the parameters or state variables. For noisier systems, we will no doubt be fitting to the noise of the system and will not obtain sufficiently smooth estimates. If $\boldsymbol{R}$ is too large, then slow convergence of the parameters can cause divergence if $\sigma > 0$. In the previous section, Figure 97 highlights an instance where many ensemble members have a positive value of σ. Despite this we still have convergence of the method. We can attribute this partly to being fortunate, but also selecting an appropriate $\boldsymbol{R}$ such that the parameters and state variables are corrected quickly enough to prevent eventual divergence.

To explore the convergence of the parameters, we choose $\boldsymbol{R}$ to be constructed as a diagonal matrix of elements δ^2_{est} with entries $\delta_{\text{est}} \in [0.001, 0.01, 0.1, 0.5, 1, 2, 10]$ to span a range of under and over-estimations of the measurement noise. We then compute the parameters after $T = 20$ to estimate convergence of the method over 200 different random seeds. This amounts to a different selection of initial conditions and different perturbations to the observations. For each value selected for $\boldsymbol{R}$ the perturbation to the observations will be the same for a given initial condition. Initially, we set $\boldsymbol{P} = 2 * \mathbb{I}$.

For each value of δ_{est} we show the collective convergence of the parameters of the different random initialisations using boxplots. Boxplots are constructed by partitioning the data evenly into quartiles. The first quartile, Q_1, is where 25% of data is below this point, the second quartile range, Q_2, is defined where 50% of data lies below this point and finally the third quartile range, Q_3 has 75% of data lying below this point. The upper and lower bounds of each box span from the first, Q_1 to the third, Q_3, quartile of the data. The inter-quartile range (IQR) is then defined as the difference between the third and first quartiles (the top and bottom of the boxes). The lower whisker on the boxplot extends from $Q_1 - 1.5 * \text{IQR}$ and upper whisker extends from $Q_3 + 1.5 * \text{IQR}$. Data outside of this range is marked as fliers by circles. Finally the median is marked by orange lines in the box.

The convergence of the parameters is shown as a set of boxplots for varying value of δ_{est} in Figure 99. Note that some outliers exist beyond the limits of the plot, but are not shown for the sake of clarity. We can see for both μ and σ, when the measurement noise is grossly underestimated we obtain poor convergence of the parameters. For values of $\delta_{\text{est}} \in [0.1, 0.5, 1, 2]$ centred around the true value, the estimates in general show better convergence. In this case the majority of the results lie in such a small IQR around the true value that the boxplot becomes very narrow. For overestimations of the noise, the convergence of the parameters again degrades. EnKF appears to have better success identifying a valid range for σ, however, we have to be aware that results with positively valued σ will likely diverge.

The convergence of the runs for different initial μ and σ are shown in Figure 100. There is no clear trend signifying which initial picks will diverge, and a large number of these initial picks diverge irrespective of δ_{est}. However, for estimates of δ_{est} closer to the true values, most initial conditions with $\sigma > 0$ can be seen to diverge. For larger values of $\delta_{\text{est}} = 10$, we can see even more initial conditions diverge. Increased uncertainty in the measurements prevents correction of $\dot{a}$ towards the observations and results in filter divergence.

Reducing the initial uncertainty to $\boldsymbol{P} = \mathbb{I}$ reduces the number of diverging runs and as shown in Figure 101. It is also more evident that divergent runs tend to start within positive initial guesses for σ. In many ways, as we do not wish for positively valued σ, it does not make sense to use this as an initial guess. Here we have included it to demonstrate that convergence is still possible regardless.

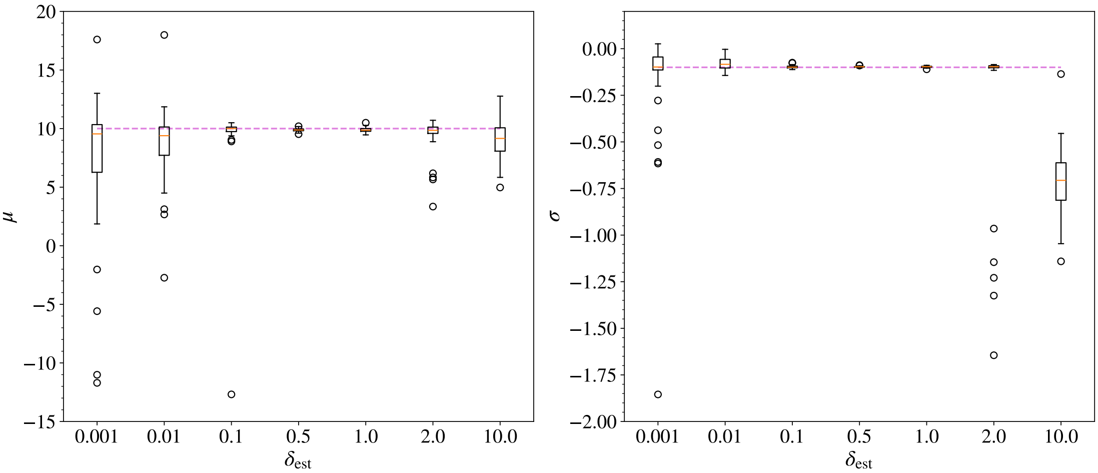


Figure 99: Boxplots for the convergence of μ and σ over 200 randomly seeded runs for different estimates of measurement noise with $P = 2 * \mathbb{I}$. Only converged results are considered, not all 200 seeds converge. The dashed magenta line shows the true values and the orange line gives the median.

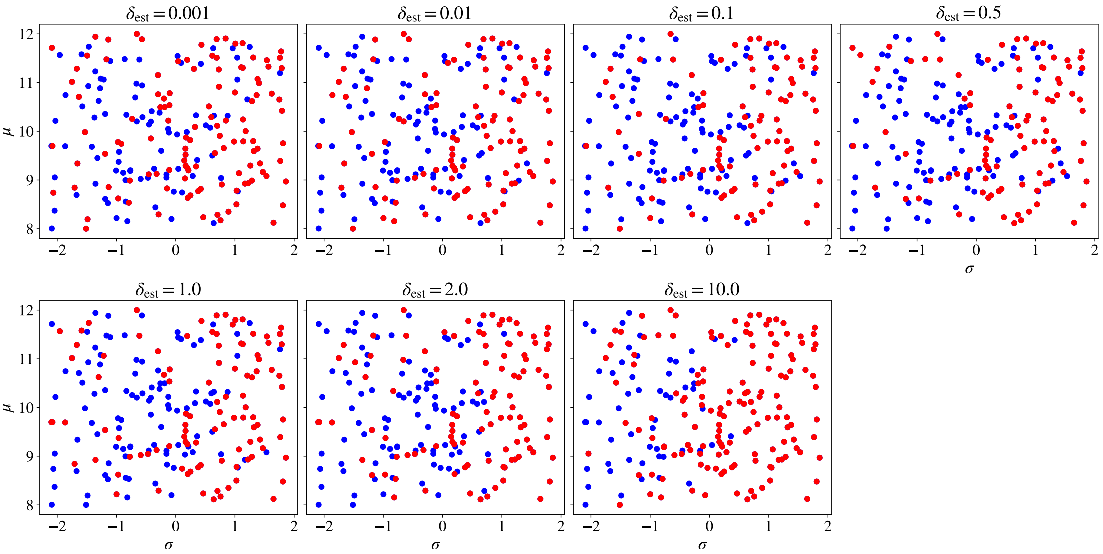


Figure 100: Scatter plot for successful runs (blue) and runs which diverge (red) for varying estimated observational noise δ_{est} and initial uncertainty $P = 2 * \mathbb{I}$.

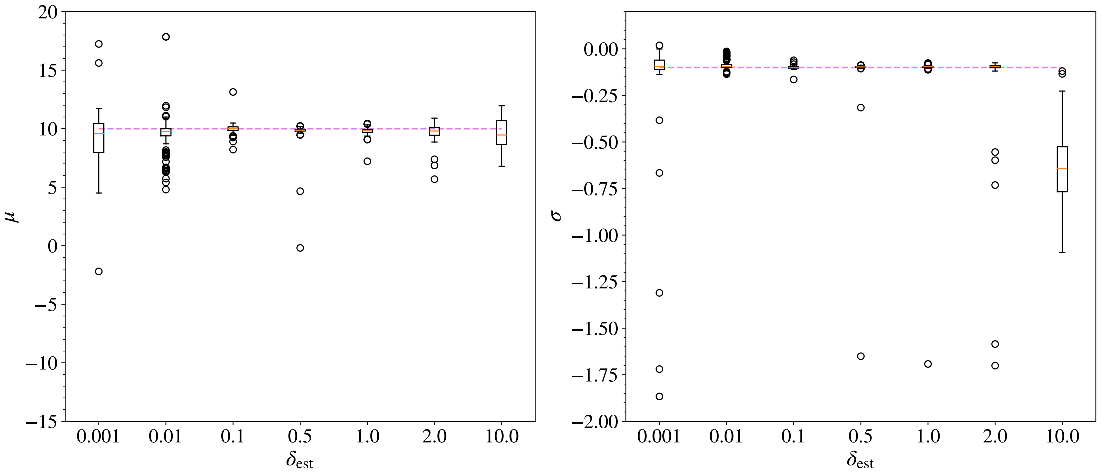


Figure 101: Boxplots for the convergence of μ and σ over 200 randomly seeded runs for different estimates of measurement noise with $P = \mathbb{I}$. Only converged results are considered, not all 200 seeds converge. The dashed magenta line shows the true values and the yellow line gives the median.

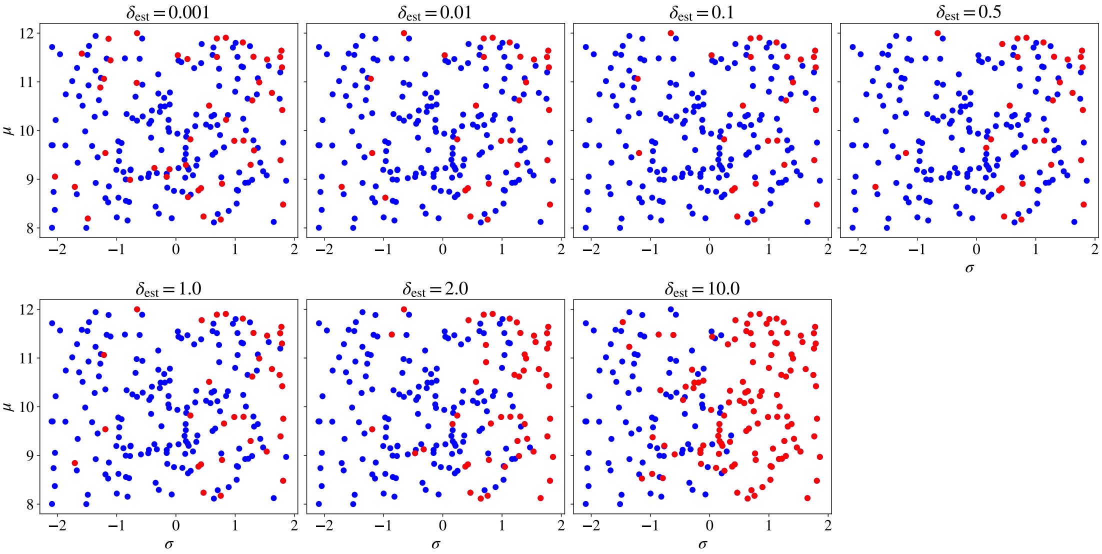


Figure 102: Scatter plot for successful runs (blue) and runs which diverge (red) for varying δ_{est} with $P = \mathbb{I}$.

6.2.11 Constrained EnKF

A general issue encountered when attempting to sequentially fit parameters of the ANAC model is the tendency of σ to take non-physical values ($\sigma > 0$). The resulting model is then unconditionally unstable and tends to infinity for all initial conditions. A further complication that has been demonstrated is that even if the ensemble mean does not violate this condition, some of the members of the distribution may still result in the routine becoming unstable within assimilation time-steps. This is further compounded when measurement noise and other factors are incorrectly estimated. One solution to this is to constrain the ensemble members to obey known physical properties.

When constraining EnKF, there are two main approaches to implementing constraints, each with their own variants [126]. The first option is to directly constrain the solution for the calculation of the Kalman gain for each ensemble member. The clear benefits of this are a solution which respects both the results of an optimal Kalman filter and the constraints, importantly still attempting to fit the observations at assimilation time-steps. The second approach is to initially perform an unconstrained EnKF estimate and then constrain the analysis members. This has the benefit of reduced computational complexity and ease of implementation, i.e., no changes need to be made to the core EnKF routine.

The first approach, implemented by [65], involves rewriting the minimisation scheme as

$$\begin{aligned} \boldsymbol{v}_c^a &= \arg\min_{\boldsymbol{v}^a}((\boldsymbol{v}^f - \boldsymbol{v}^a)^T(\boldsymbol{C}^f)^{-1}(\boldsymbol{v}^f - \boldsymbol{v}^a) + (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^a)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}^a)), \\ &\text{s.t. } \boldsymbol{v}_{lb} \leq \boldsymbol{v}^a \leq \boldsymbol{v}_{ub} \end{aligned} \tag{6.58}$$

where $\boldsymbol{v}_{lb}$ and $\boldsymbol{v}_{ub}$ are the upper and lower bounds respectively, while the calculation of all other variables remains the same. The state variable $\boldsymbol{v}_c^a$ denotes the now constrained result. When performed on the prior ensemble members, this now effectively becomes the update step in place of the Kalman gain. This is not a common approach in EnKF, as it involves solving this system for each ensemble member. An alternate viewpoint is to perform the unconstrained update of the state and then solve the system ([126])

$$\begin{aligned}&\arg\min_{\boldsymbol{v}_c^a}((\boldsymbol{v}_c^a - \boldsymbol{v}^a)^T(\boldsymbol{C}^a)^{-1}(\boldsymbol{v}_C^a - \boldsymbol{v}^a) + (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}_c^a)^T\boldsymbol{\Gamma}^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{v}_c^a)),\\ &\text{s.t. } \boldsymbol{v}_{lb} \leq \boldsymbol{v}_c^a \leq \boldsymbol{v}_{ub}\end{aligned} \tag{6.59}$$

purely on the analysis states. Other authors sometimes only minimise [75]

$$\arg\min_{\boldsymbol{v}_c^a}((\boldsymbol{v}_c^a - \boldsymbol{v}^a)^T(\boldsymbol{v}_c^a - \boldsymbol{v}^a)) \qquad \text{s.t } \boldsymbol{v}_{lb} \leq \boldsymbol{v}_c^a \leq \boldsymbol{v}_{ub} \tag{6.60}$$

assuming that the fitting to observations has already been performed in the unconstrained step. Here we will mainly focus on constraints applied to the analysis states due to their simplicity.

When considering constraints applied to EnKF, we must also clarify applications of constraints to the ensemble members individually and constraints applied to the means and covariances. One obvious issue is that constraints applied to the entire distribution do not then necessarily apply to the mean with the opposite situation also being the case. Nonlinear constraints applied to Kalman filtering are discussed by [59], where it's concluded that constraints applied to the ensemble members only carry over to the mean if the applied equality constraints are linear. Thus in some cases one must consider both constraints on the distribution and the mean. A general discussion of these constraints without application specific context has been given by [118] more recently. Here we largely consider constraints outlined by [75] and while being application specific, these are relatively easy to generalise.

In total, three different constraint approaches are outlined by [75]

1. The naive constraint method: Ensemble members are simply set to upper or lower bounds of the constraints if they are violated. The explicit implementation of this is discussed by [118].

2. The accept/reject method: If ensemble members in either the forecast and analysis (or both) distributions violate constraints the respective model or measurement errors are reinitialised with a different seed and the members redrawn until the constraints are satisfied [61]. This has an obvious downside in that there may be no draw which satisfies the constraints if initial guesses are poor.

3. A projection based method: the unconstrained estimated states are projected to a new set of states which minimises the mean-squared error between new states which satisfy the constraints, and the previous unconstrained estimate of the posterior [39]. In this sense we attempt to retain the original solution where possible.

All the methods suggested above apply constraints to the estimates of the posterior and do not directly constrain the calculation of the Kalman gain. The accept/reject method can become cumbersome to implement if model error (additive noise) is implemented on integration time-steps. This would involve many more checks per assimilation step and depends highly on finding draws which satisfy the constraints. The naive constraint method presents an issue that setting members to some upper or lower bound produces truncated distributions which clearly will not be represented by a Gaussian. This also weights the calculation of the means to the bounds which is not necessarily desirable and will depend strongly on the mean of the prior. While success is suggested for these methods by [75], no direct comparison to an unconstrained situation is made and so it is not clear if the final convergence of the parameters is any better as a result of the constraints. We would suggest that an important consideration in approaches which modify the ensemble members directly is that they can artificially reduce the ensemble spread and thus underestimate system uncertainty. In our case, they would principally aid in stopping divergence of the filter.

Finally, other work does exist for constraining EnKF such as the work by [126]. In this publication, constraints are again applied to the estimate of the posterior through an optimization of the predicted distribution to the physical one by minimising the Kullback-Leibler divergence. Given that the Kullback-Leibler divergence measures the similarity in two multivariate distributions, this process results in a new reshaped estimate of the posterior distribution which better fits the constraints. A more extensive comparison of the constrained to unconstrained methods is presented in this work and may be worth considering in the future as it maintains some spread in the ensemble members.

6.2.12 ANAC with constraints

We test the naive constraint method where ensemble members are simply reset to the upper or lower bounds of an interval as an alternative approach to optimising for initial conditions. For the ANAC model we must set $\sigma < 0$ always which involves constraining each ensemble member such that this condition is true. This must be applied at multiple different instances: when a forecast is made, when noise is added and when the analysis states are found.

We then repeat the above analysis in the single measurement case for perturbations in $[-2, 2]$ from the true values. Again this will result in initial guesses with $\sigma > 0$ and does not particularly make sense when we hope to constrain $\sigma < 0$. It is included so that the case of many ensemble members being constrained can be tested as we expect these cases to perform poorly due to a large artificial reduction in uncertainty. If many members are constrained to the upper bound of $\sigma_{ub} = -1 \times 10^{-16}$, the spread of the ensemble members will be small (implying the uncertainty is small) and we will effectively limit ourselves to a class of simple harmonic oscillators where only μ is varying and the effect of the a^3 term is negligible.

Figure 104 shows the results of the parameter estimation using the simple constraining procedure for varying estimates of noise. Compared to Figure 101 there appear to be many more cases of μ diverging from the true value even for correct estimates of the noise. Similarly for the estimates of σ there are many more estimates lying close to 0. This is hardly surprising, given that the distributions are artificially truncated at the bounds. Further, after constraining, no initial guesses become unstable and hence shows all filter divergence previously can be attributed to $\sigma > 0$ either by some of the ensemble members or the mean. As the ANAC model is energy-preserving for $\sigma < 0$, we should expect stable solutions even if the resulting estimates are poor.

An instance of poor parameter convergence is given in Figure 103. We can see that the constraints cause a rapid collapse of uncertainty in σ at the first assimilation step with the mean centred close to the upper bound of the constraining interval. The confidence intervals then no longer span the true values and the routine is unable to correct to a model which has a non-zero cubic nonlinearity. It is important to realise though that the confidence intervals for sigma are not good estimates of the covariance as we assume the distribution can be representing by a Gaussian, however, this is clearly not the case with constraints. In general in cases where many ensemble members are constrained we can obtain poor convergence as the mean is weighted to the bounds. This can easily be corrected by including a small degree of process noise for σ, maintaining some ensemble spread after it has been constrained.

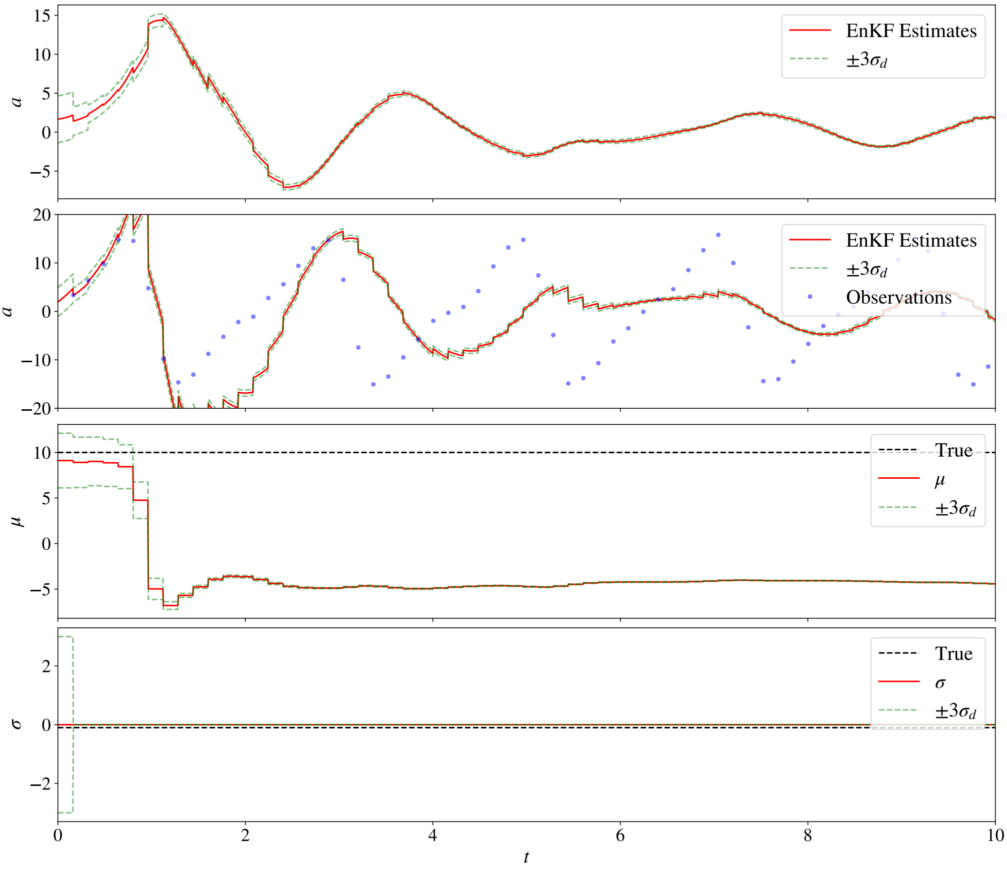


Figure 103: Case of poor parameter convergence when using constraints. A large initial uncertainty is given to σ which causes a collapse of ensemble members at the lower bound.

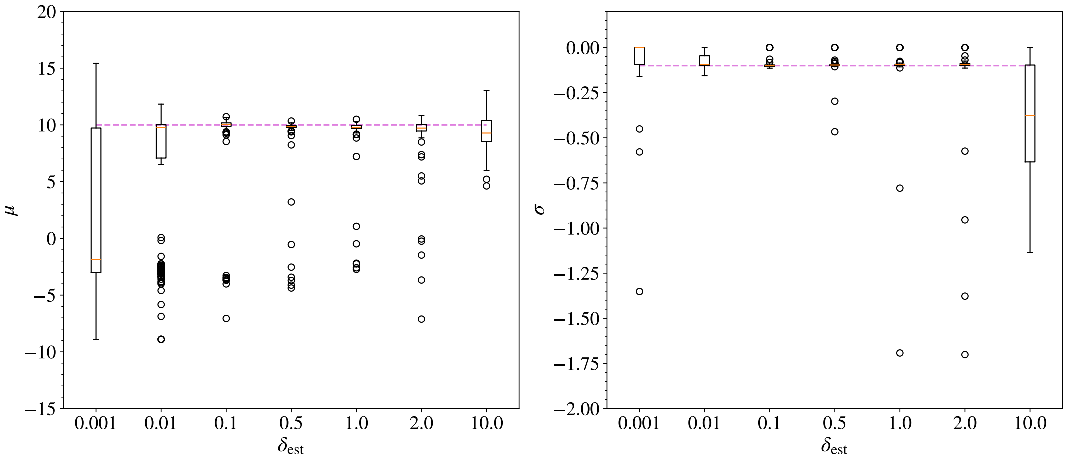


Figure 104: Boxplots for the constrained parameter estimation over 200 seeded runs for varying estimates of measurement noise. The left figure shows convergence of μ and the right figure shows convergence of σ.

Solving the constrained optimisation is substantially faster than solving an optimisation problem over the first oscillations of a system. However, serious care needs to be taken that the results are converged and the initial guesses used make sense.

6.3 Concluding Remarks

In this section we have introduced the theory to EnKF and outlined several important extensions and additions which have been implemented. EnKF is a Monte-Carlo estimation method which uses an ensemble of members to represent the state and uncertainty. By making Gaussian assumptions about the spread of the ensemble members, we showed how several simplifying arguments can be made. EnKF then essentially reduces to an iterative update model which corrects each ensemble member based on the calculation of the Kalman gain. The new statistics are then calculated from an equal weighting of all the particles.

While EnKF code has been used already in similar contexts (see ref. [115]), we highlighted a number of issues with this code which made it challenging to run. As a result, existing code for EnKF from the `filterpy` package was adapted for use here. This code was extended to include constraints on the ensemble member distributions and the specification of inflation factors.

Application of this code was performed to validate its performance in several different scenarios with the ANAC model derived in § 2.4.2. The first of these considered cases where both measurements of $\dot{a}$ and a were available. We also addressed concerns that incorrect parameter guesses and initial conditions required an additional optimisation process given in ref. [106]. In the examples considered here, we found that the convergence of the method was very good in sparsely sampled noisy signals with the same initial perturbations applied to the initial parameter estimates given in [106]. Instead we found that the main concerns on robust parameter convergence come from: 1. how the method is initialised and 2. if certain regions of the ensemble space can generate unstable models. In the former, it is important to consider sensible initialisation of EnKF which may involve relying on physical intuition of the parameter space. The general importance of specifying sensible initial priors is also discussed which is remarked in [55] but surprisingly not mentioned by [106, 115]. It is unclear if this has impacted the results of this work, but it is likely sensible priors were used without mention. For the latter concern, we added a constraint method which can be used to keep ensemble members in a sensible region of the potential parameter space.

In the final sections, we began to progress application to more experimental-relevant conditions in which only partial measurements of the state are available. This means complete inference of all the parameters must be made with only one observation. EnKF still managed this situation well, and it even robust to errors in the estimated measurement uncertainty. Even in cases where the measurement uncertainty is incorrectly estimated, incorrect estimations can be inferred from the confidence intervals generated by EnKF.

In the following chapter we will consider the important extension of this work to the ANAET model. This model is more relevant in experimental conditions and we will consider how EnKF performs in these cases.

7 EnKF and the ANAET model - progressing to experimental conditions

In this section we implement EnKF with the ANAET model. We will pay particular attention to the performance of EnKF with respect to the following conditions i) partial observations of the state variables ii) robust to observational noise iii) errors resulting from initial parameter estimates. We will conclude this section by introducing cases with non-zero process noise which are expected to be essential when applying to experiment.

7.1 EnKF and the ANAET model

We now extend the analysis to include the ANAET model which has the form

$$\dot{a} = v, \tag{7.1}$$

$$\dot{v} = -\gamma_r a - (\mu_1 + \mu_2 b)a^3 - \mu_6 a^6 v, \tag{7.2}$$

$$\dot{b} = \nu_1 - \nu_2 b^2 - (\delta_0 + \delta_1 b)a^2. \tag{7.3}$$

where $\gamma_r, \mu_1, \mu_2, \mu_6, \nu_1, \nu_2, \delta_0$ and δ_1 are all constant real-valued parameters. The extension from the ANAC model to the ANAET model is important when matching to experimental data as discussed in § 2.4.3. However the most salient points in this section are the inclusion of an equilibrium mode b and the two extra terms. The first of these is $-\mu_2 ba^3$ where the multiplication of b acts to destroy and recreate the potential well. The second is high-order diffusion term $-\mu_6 a^6 v$ which reduces the amplitude of $\dot{a}$ when it grows large causing a crash of the b mode. In relation to experiment, this creates a signal resembling a sawtooth signal seen in electron temperature measurements and a gong signal from magnetic measurements. We are therefore interested if this extended model can be fit using EnKF.

The first challenge in fitting this model is the increase of the number of state variables and free parameters which must be determined. Extrapolation of arguments in the previous section suggests that normal distributions representing all parameters presents more opportunities for ensemble members to lie in unstable regions of phase space. In relation to experiment, we must eventually consider cases of measurements of only $\dot{a}$ and b, which means we must infer many parameters and an unknown state variable from incomplete observations. The second challenge results from the multiscale nature of the ANAET model. The oscillations of $\dot{a}$ (equivalently v) are on a much faster timescale than the equilibrium mode b, a feature which is also typically seen in experiment (see for example [17]).

In this chapter we will present a progression of work which first implements the ANAET model with EnKF and assesses the general challenges of doing so. We will then restrict to cases with partial measurements, attempting to assess how robust EnKF is when only one or two observed state variables are available. The case of a single state measurement could be relevant if we wish to only fit a single diagnostic, for example. We will also assess the sensitivity of EnKF at different sampling rates and degrees of noise so that we can anticipate under what conditions it will function. Finally we discuss extensions to experiment relevant conditions where either both the noise is high and sampling rate is low, or the observations do not display regular sawtoothing behaviour. Both of these are expected challenges when addressing experimental data.

7.1.1 Computational challenges of the ANAET model

The ANAET model is a stiff problem and requires an integrator which can switch between stiff and non-stiff solvers. For this we make use of scipy's LSODA solver which essentially is wrapped code for the fortran solver ODEPACK from [12]. One limitation to this approach is due to the class structure of the ENKF solver. At each integration time step, the class calls the function as in Listing 5 which involves invoking the Python interpreter. If the integration range in `t_span` (between assimilation time-steps) is large enough, this overhead is not particularly noticeable. However, if we call this function recursively at integration time-steps, the solver slows substantially. This is an issue when additive noise or process noise is being used, and also when assimilation points are frequent.

```
observations =  solve_ivp(ANAET,
                          t_span=t_span,
                          y0 = x0,
                          t_eval=observation_times,
                          method="LSODA",
                          rtol=1e-6,
                          atol=1e-8, vectorized=False).y.T
```

Listing 5: Function call for scipy's LSODA method.

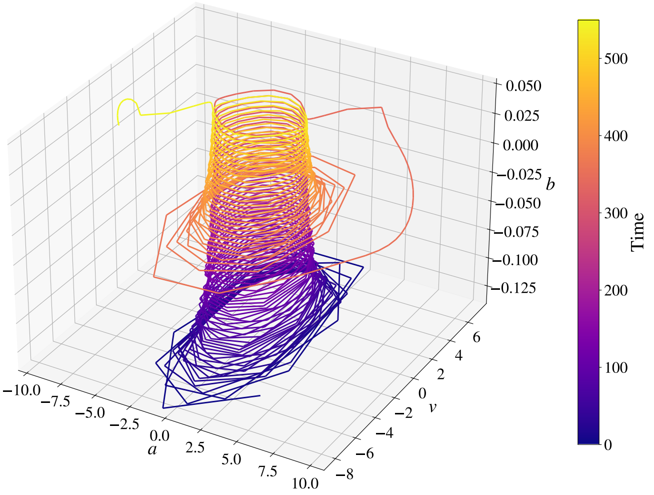


Figure 105: Phase space for the training data of the ANAET model.

To mitigate this, we implement two different approaches

1. Integration of the ensemble members between assimilation time-steps to reduce the number of function calls. The priors are then calculated after integration rather than concurrently which can further speed code and calculation of covariance matrices is only necessary on assimilation time-steps. If assimilation time-steps are small this will not provide a great speed up. Further, depending on how process noise is applied this may not be suitable.

2. Implementation of an alternative function call with `numbalsoda` [169] which avoids invoking the Python interpreter. This can be used to retain the standard class structure, or in the same way as scipy to integrate between assimilation time-steps. Integration within the class structure can still be slow though as vectorisation of these calculations is faster. `numbalsoda` also allows for `numba` compiled functions to be used.

In general, this can still be a stiff problem to solve which can be slow at particular parameter values. Using `numbalsoda` [169] and integrating between assimilation time-steps yields the fastest performance for this problem. Approximate timings from single runs are shown in Table 5 to gain an idea of where the overhead exists. Note that these are not benchmarked timings and are only representative of example times. From this we can see that calling `solve_ivp` LSODA at integration time-steps is extremely slow, generating a substantial overhead from invoking the Python interpreter. This usually makes the code a factor of 10 times slower than the `numbalsoda` counterpart. The remaining speed-up can be associated with the calculation of the means and covariance matrices in a vectorised form vs within a recursive call of the class. For these cases, we calculate the mean of the ensembles at all integration time-steps, but the covariance matrices are only calculated at assimilation time-steps. While calculation of covariance matrices can be done at integration time-steps, this is a more-expensive operation as it involves matrix multiplication. The covariance matrices are not strictly required outside of assimilation time-steps. For this example 40 ensemble members are used so we should expect the ENKF process to run more than 40 times slower in serial.

Method	solve ivp	solve ivp t_{assim}	numbalsoda	numbalsoda t_{assim}	observational
Time	∼18 mins	∼11s	∼1 min	∼0.8s	0.04s

Table 5: Table of representative times for each solving method for the time interval $t \in [0, 120]$. The final column lists the time taken to create the entire observational data-set.

Further speed-up could be achieved by either parallelising over ensemble members or vectorising the integration of ensemble members. Both of these options are not particularly viable in Python due to the overhead in CPU parallelisation and strict typing of `numbalsoda`. Other options include parallelising the calculation of the covariance matrices with `numba` but this will likely give minor speed-ups as the covariance matrices are typically small in our case. The main computational expense comes from evaluating the space-state model. A preferable point of speed up would be to avoid using an external integration package entirely and make use of user defined integration methods within Python which could be parallelised. Attempts to do this again resulted in minor improvements over using `numbalsoda` alone, highlighting the general poor performance of Python as a parallelisable language.

In terms of parallelising over ensemble members, we could perform integration of ensemble members in parallel which should work well. Some issues with this though are held within the class definition of EnKF. `joblib` requires that the parallelised objects can be pickled, which is not possible with classes so it could only be used within the class if user specified solvers are implemented. `multiprocessing` requires `if __name__ == __main__` to run correctly and so cannot be written within the class definition which is the most logical place to include it. `Dask` may be able to solve some of these problems however it has not been possible to develop the code to test this within the remaining time.

7.1.2 Analysis of the ANAET model with ENKF

	a_0	v_0	b_0	γ_r	μ_1	μ_2	μ_6	ν_1	ν_2	δ_0	δ_1
Value	-1.6	-0.0099	0.0086	1	0	-2	1×10^{-4}	0.001	0.005	1×10^{-4}	0

Table 6: Initial conditions used with the ENKF method for the ANAET model.

The ENKF method in all subsequent sections is applied using the parameters and initial conditions in Table 6 unless otherwise stated. These initial conditions are chosen where the ANAET model shows sawtoothing behaviour. The noise for (a, v, b) is chosen to be 10% of the standard deviation of each respective variable as this is the degree of noise expected from experiment [111]. This corresponds to a selection of $R_a \approx 0.3^2$, $R_v = 0.25^2$, $R_b = 0.004^2$. Here b varies on a smaller scale than the other signals and must have a smaller uncertainty compared to the other variables. For the parameters, we choose the initial uncertainty to have a standard deviation of ±10% of the parameter value. In general the approach of validation and benchmarking with covariance diagonal elements being proportional to the magnitude of true values in not an uncommon assumption [75]. Unless otherwise stated, when the parameter is expected to be zero, the initial uncertainty is set to zero which prevents EnKF altering the parameter value.

Previously with the ANAC model, sampling rates were based on characteristic timescales from the FFT of the a equation (equivalently $\dot{a}$) equation. In this instance, basing sampling rates from the FFT is more challenging as the problem spans multiple long and short timescales. As the mean-field component of b grows larger, the oscillations in a and v slow until this leads to a sudden reduction in b after which a and v oscillate on a faster timescale. As such we characterise the sampling rate from the faster oscillations for which we obtain an approximate period of 3.3. Experimentally speaking, observations of the core temperature (b) are concurrent with the Mirnov measurements (a) so classifying the sampling rate on the fastest time-scale is a reasonable assumption. We choose the sampling rate for the fast oscillations initially to be $\nu_{\mathrm{f}} \approx 6$ giving $t_{\mathrm{assim}} = 0.55$. As the period of the system is not fixed, we will typically refer to the assimilation time-step rather than the sampling rate in later cases.

7.1.3 Case 1: Full-state observations

$\boldsymbol{R}$	$\boldsymbol{Q}$	$\boldsymbol{P}$	δ_{obs}	N
$\delta_{\mathrm{obs}}^2 * \mathbb{I}$	$\mathbf{0}$	10% of parameter value	10% of signal sd	100

Table 7: EnKF class parameters used for the assimilation of the ANAET model.

We first start by remarking that in the case of no observational noise, we achieve excellent convergence of the parameters even when initial guesses are incorrect in excess of ±10% of their true values. However, this is only in the case where the measurement noise is correctly estimated at $\boldsymbol{R} = \mathbf{0}$. If this is not the case, we obtain extremely poor estimates despite accurate initial guesses. This can be mitigated by including process noise to offset the impact of artificially perturbing the measurements. As this is a trivial case, the results are not shown. If we perform this process with a small inflation factor, no appreciable gain in parameter convergence is found.

We next consider the case of full-state measurements (explicitly a, v, and b observable), perturbing parameters by a standard deviation of 10% of their magnitude and perturbations with standard deviation $[1, 1, 0.1]$ for each state variable. The initial conditions and uncertainties are chosen as described in the previous section with the random seed set to 99. We use a smaller uncertainty in b as this evolves on a smaller scale compared to a and v. In the case of full-state measurements, we can be relatively certain of our initial estimates for a, v, b. The settings for the EnKF class are shown in Table 7 where initially we use correct evaluations of the measurement noise. We use a larger number of ensemble members here at $N = 100$. For lower numbers of ensemble members parameters tend to converge quickly to poorer estimates suggesting that we suffer from sampling based errors. Given that the dimension of the problem is now up to size 11, it is reasonable to expect an increase in the number of ensemble members to accurately capture the correct dependencies.

Speed-up could no doubt be obtained by exploring smaller ensemble sizes and the use of various inflation factors to counteract sampling errors. For full-state measurements, there are cases where $N = 30$ ensemble members produces promising results.

Figure 107 shows the results of the state variable convergence. We can see that the state measurements quickly converge to promising estimates for all variables. For the parameter convergence in Figure 107 we also see good parameter convergence for all parameters. From the location of the first crash in amplitude of the b mode at $t \approx 90$, we can see that γ_r, μ_2, μ_6 and δ_0 are poorly determined until this point. Following this, the confidence intervals on each of these parameters converge significantly. For μ_2, μ_6 and δ_0 these all represent coupling terms between the slow equilibrium mode b and the faster dynamics a. Convergence of γ_r will then depend heavily on the convergence of these parameters and thus converges to its true value following correction of the other parameters. Following this first crash, the uncertainty in these parameters reduces substantially and there is some stagnation away from the true values.

ν_2 is the slowest converging parameter for this model because it is determined by the slow evolution of b and requires many slow oscillations until it is fully converged. If the coupling to a is ignored in the equilibrium equation of b, we have the normal form of a fold bifurcation where the solution should grow to stable fixed point $\sqrt{\nu_1/\nu_2} \approx 0.45$ which does not happen due to the crashing behaviour described. It seems likely that only the linear growth of b is well determined before crashing occurs. Further, as the perturbation to the initial conditions is Gaussian, for this seed the true value lies further from the confidence intervals for ν_2. The same principle also applies to the convergence of ν_1. The convergence of these parameters suggests that optimising over the first few fast oscillations for better initial conditions as suggested by ref. [106] may not be beneficial because many parameters converge over multiple slow oscillations.

To test the convergence with different seeds, we fit the observations with 30 randomly selected initial conditions and compute the largest percentage errors for each parameter over all the runs with results presented in Table 8. In all runs the parameters converge well, with the largest absolute percentage error in ν_2 at 6.5%. The means over all the seeds are observed to be close to the expected value suggesting that the convergence is robust. Typically it appears that convergence of a subset of parameters is more challenging than others, and are parameters which are fixed on a slow time-scale. Two further cases are tested in Table 8, one with a multiplicative inflation factor c and the other with an additive inflation factor labelled GK of 1×10^{-8}. Here GK refers to the Gaussian kick introduced in ref. [115] which essentially amounts to an additive inflation factor. The multiplicative inflation does not appear to help improve poorer converging cases whereas the additive inflation factor could be beneficial. Given the discussion on inflation factors, there is no great expectation that they should provide marked improvement in our case.

We then repeat the analysis, allowing for δ_1 and μ_1 to take non-zero values. We perturb these parameters from a Gaussian distribution with standard deviation of 10% of each parameter value. The convergence of the parameters is shown in Figure 108. Despite adding two additional parameters to be determined with arbitrarily chosen perturbations in the parameters, we still achieve good convergence of the parameters. Both of these parameters are correctly identified as being zero after a short period. This case is relevant if we wish to consider fitting a more general model which only obeys the outlined symmetry constraints given in § 2.4.2. Finally, tests with larger perturbations in the initial conditions of one standard deviation of each parameter still have good parameter convergence, suggesting that larger uncertainties are possible. We also find that initialisation of the data assimilation along different points of the trajectory does not change the results presented.

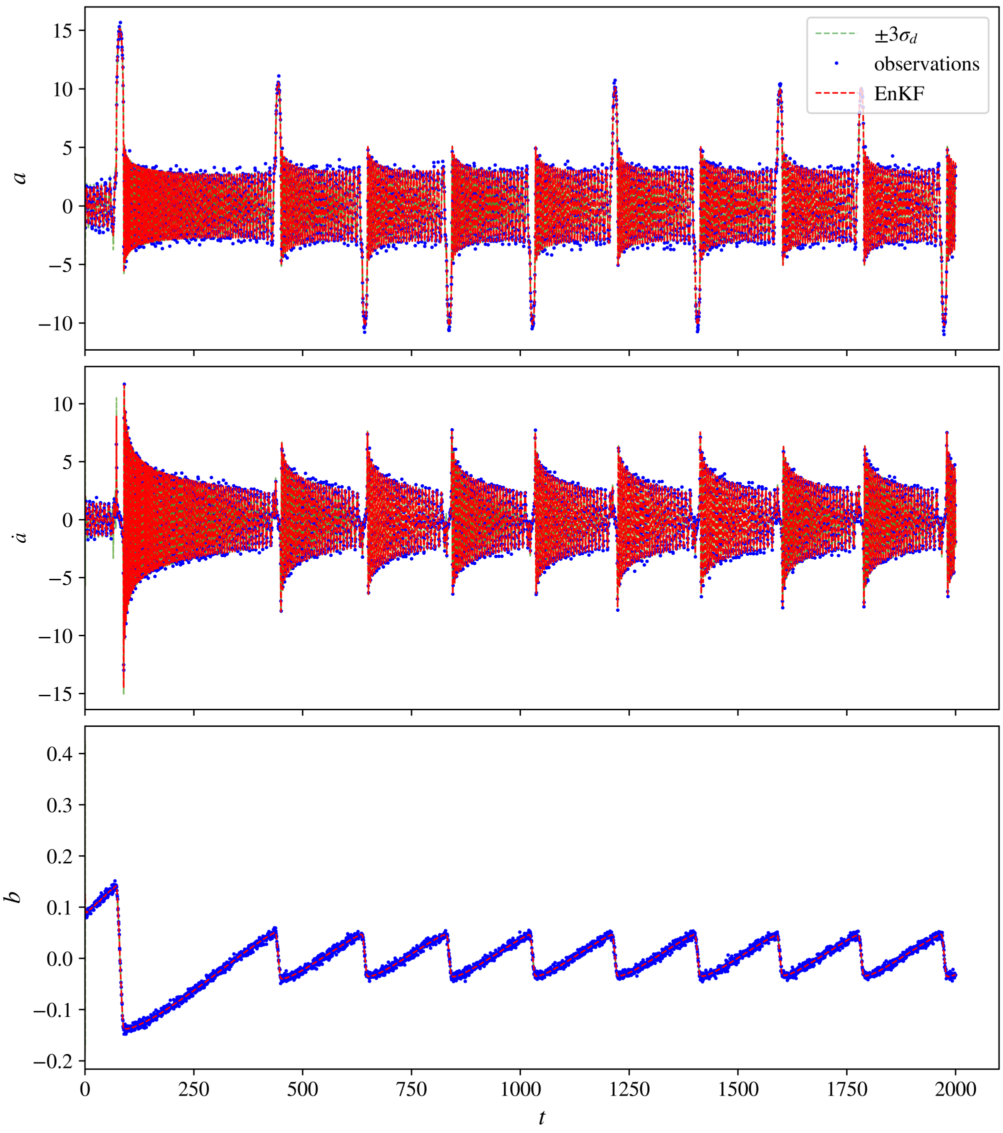


Figure 106: State convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The observational error is set to the variance of the added noise.

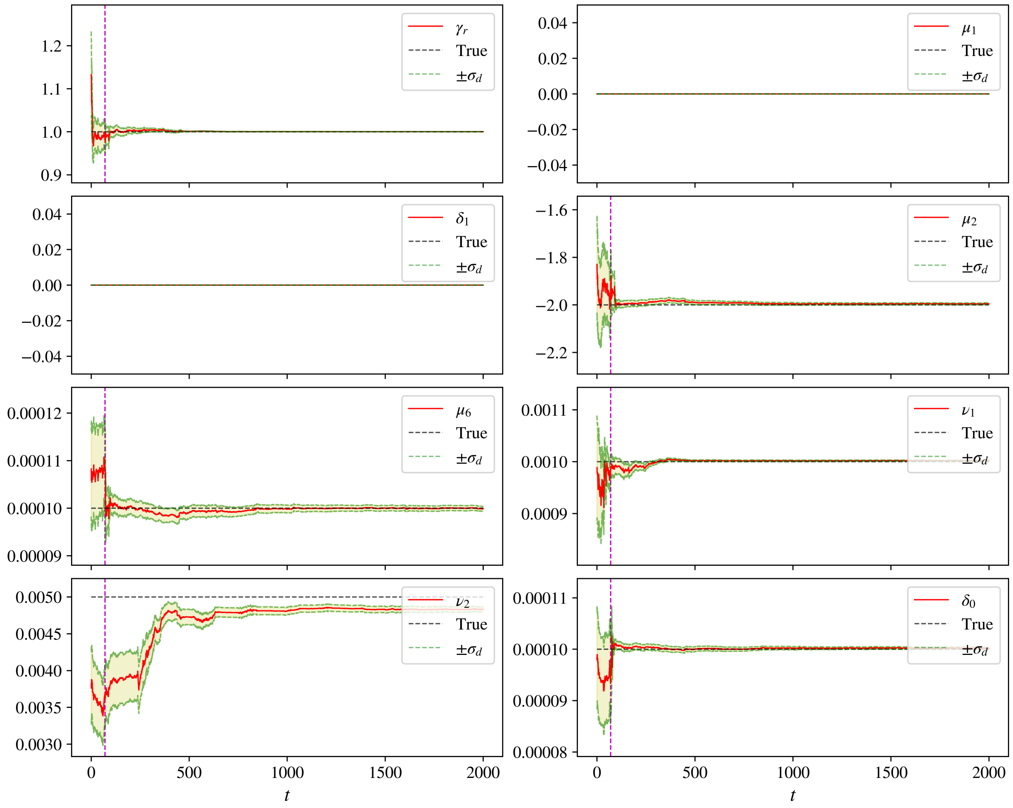


Figure 107: Parameter convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The dashed magenta line represents the approximation location of the first crash in amplitude of b. In this case only terms which appear in the model which generated the observations are allowed to take non-zero values.

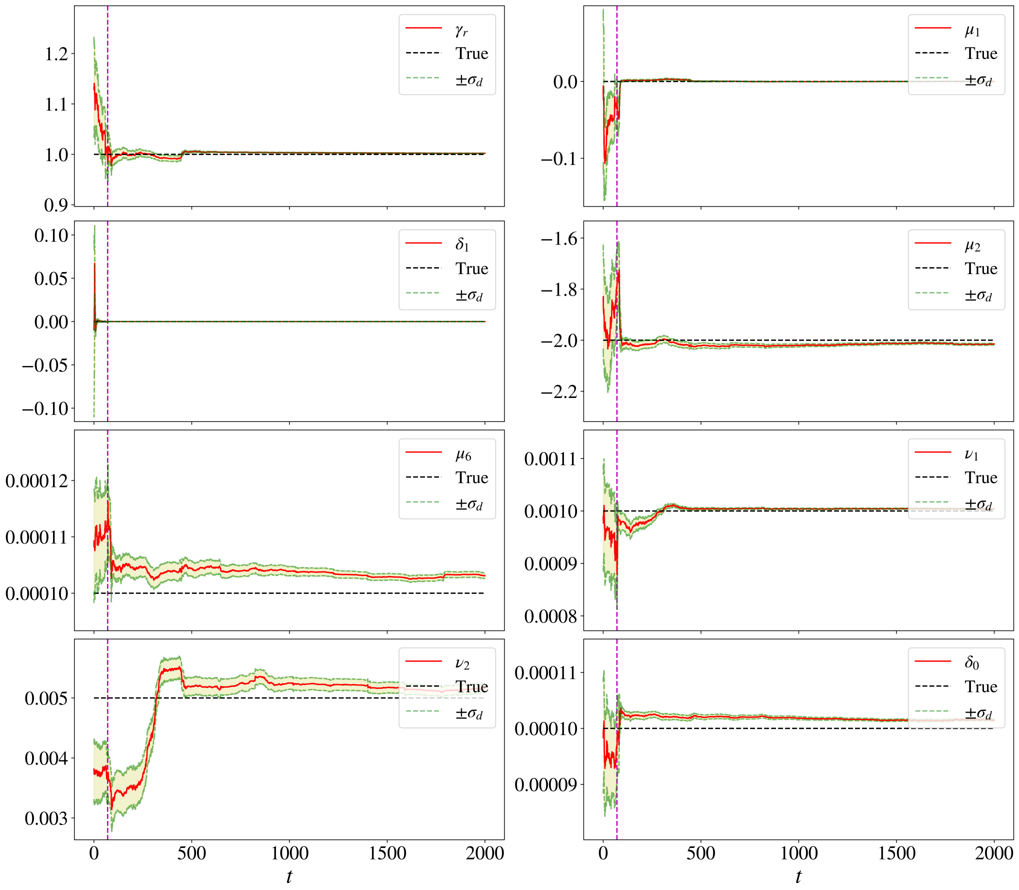


Figure 108: Parameter convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The dashed magenta line represents the approximation location of the first crash in amplitude of b. In this case all terms are allowed to be non-zero.

	c	GK	γ_r	μ_2	μ_6	ν_1	ν_2	δ_0
Largest absolute error (%)	1.0	0	0.086	1.15	3.78	0.46	6.55	1.24
Largest absolute error (%)	1.0005	0	0.086	1.29	4.05	0.41	6.47	1.28
Largest absolute error (%)	1.0	1×10^{-8}	0.063	0.95	2.9	0.433	6.40	0.90

Table 8: Largest absolute percentage errors for the different identified parameters in 30 seeded runs.

7.1.4 ANAET Single measurement systems

We now consider single measurements of the ANAET model, fitting only measurements of the a variable. We use the same initial covariances as in the previous example, with the obvious modification that $\boldsymbol{R}$ is now a 1×1 sized matrix. The results of the state variable assimilation for the single measurement system are shown in Figure 109. The observations for a are shown as blue dots and the true state for $\dot{a}$ and b, which are not observed, are shown as black dots. Note that only a section of the state variable data is plotted here for clarity. We can see despite only single measurements with incorrect initial state and parameter estimates, we can obtain good approximations of all state variables. The only exception is in b as it varies on a slow scale initially and can hence appear as approximately constant until the high order nonlinearity is non-negligible.

The parameter convergence in this example is shown in Figure 110. As can be seen, the convergence of the parameters is again much slower in the single measurement case. We can see that in all fitted variables except γ_r, convergence to the true value requires more time than the full-state measurement case. We also note that μ_2, ν_1 and δ_0 appear to be correlated in exactly the same way as one another for a large section of the data. These all represent terms in some way associated with the unstable equilibrium mode b. When only measurements of a are available, they are likely correlated in the same way. Despite slower parameter convergence, there is still a fair convergence in all parameters for this seed.

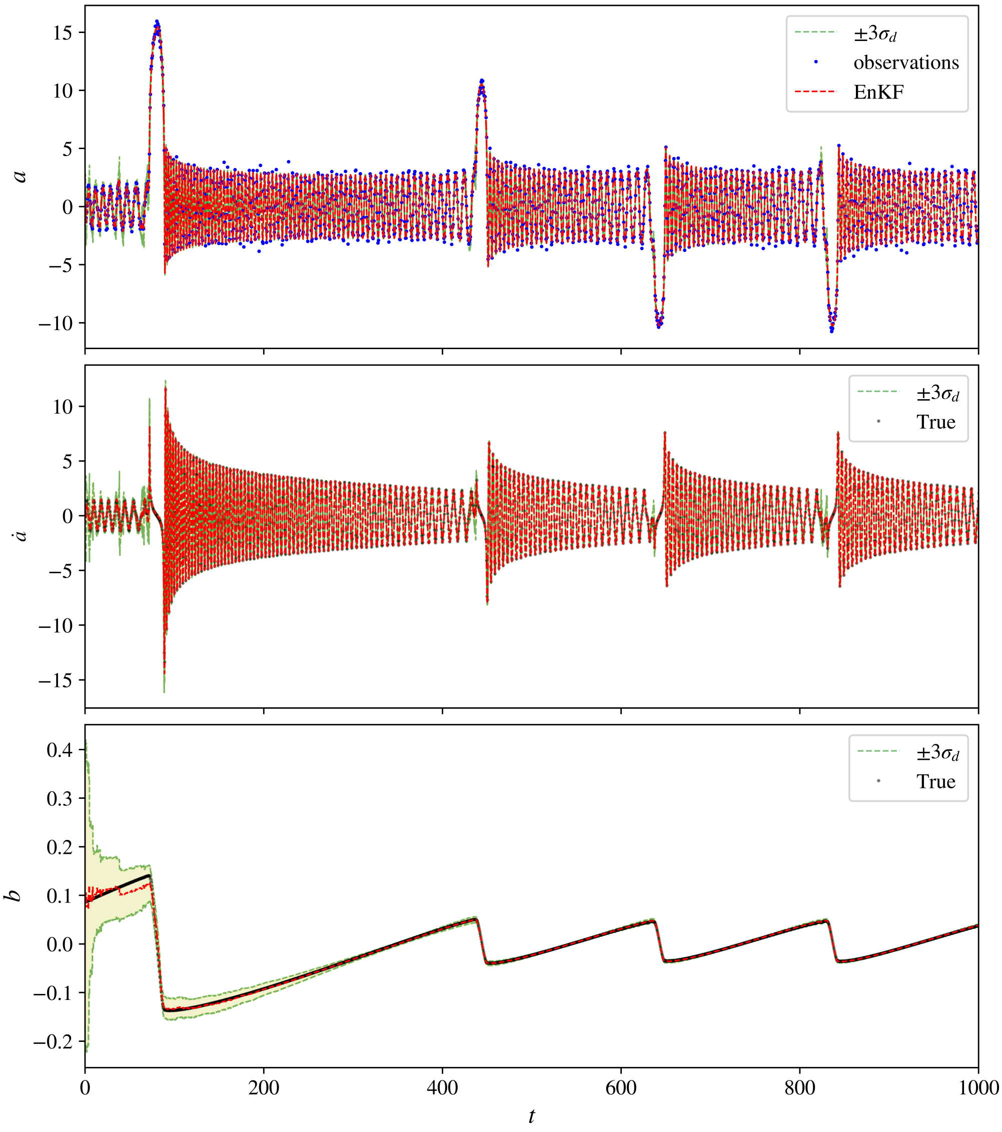


Figure 109: State convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. Only observations of a are provided here shown in blue dots.

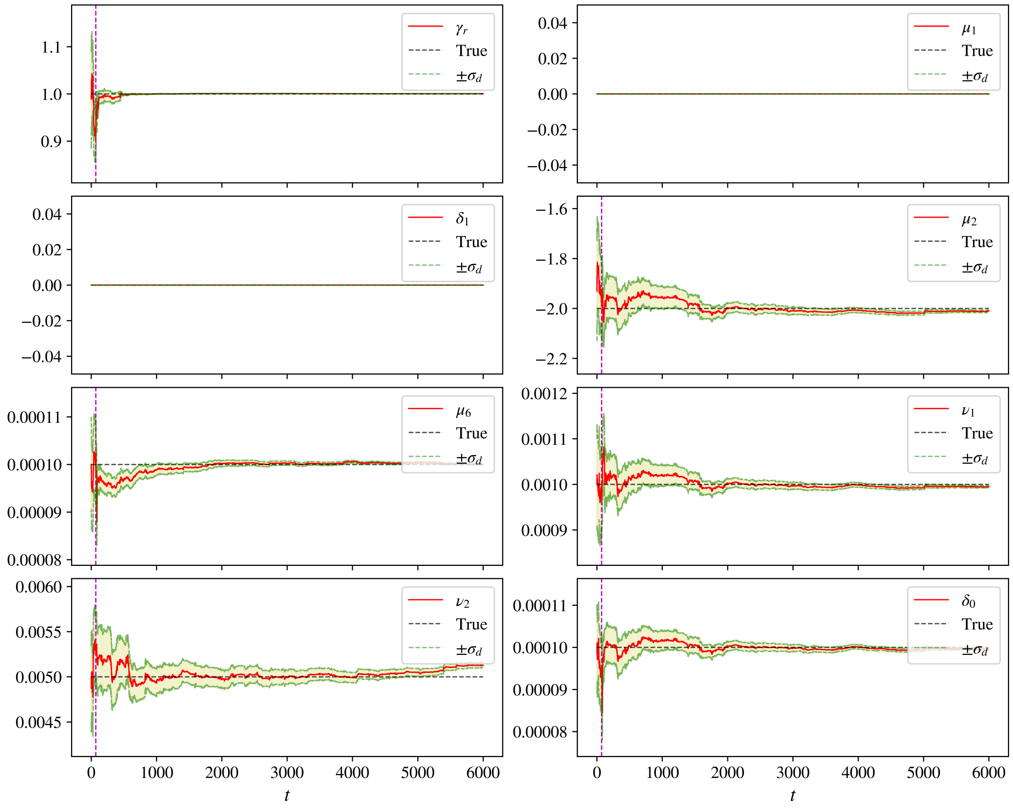


Figure 110: Parameter convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The dashed magenta line represents the approximation location of the first crash in amplitude of b. Only observations of a are used in assimilation.

We again perform a similar analysis in the single measurement case where convergence of the parameters is assessed over 30 randomly seeded trials. The convergence in this instance is measured after a longer time-period at $t = 6000$. The results are shown in Table 9. We can see that for all parameters associated to the growth of the slow mode, the maximum observed error is much higher despite a longer time for convergence. As the perturbation to the true initial condition can result in initial guesses which lie outwith $\pm\sigma_d$ of the true parameters, we can obtain poor convergence. Conversely, γ_r and μ_6 have strong correlations with a and therefore have quite robust identification.

The poor convergence for one seed is shown in Figure 111. We see that the selection of initial conditions of parameters associated to the slow mode all lie outwith $\pm\sigma_d$ and converge to local solutions for the problem. The evolution of the state variables in Figure 112 shows that despite incorrect growth of b, we still obtain the correct behaviour for a. If we consider a scaling of b which attempts to leave the dynamics invariant, we can write $b = 1/\alpha\bar{b}$ where $\alpha \in \mathbb{R}_{>0}$ we can write

$$\dot{a} = v, \tag{7.4}$$

$$\dot{v} = -\gamma_r a - (\mu_1 + \tfrac{\mu_2}{\alpha}\bar{b})a^3 - \mu_6 a^6 v, \tag{7.5}$$

$$\dot{\bar{b}} = \alpha\nu_1 - \tfrac{\nu_2}{\alpha}\bar{b}^2 + (\alpha\delta_0 + \delta_1\bar{b})a^2. \tag{7.6}$$

Thus we may expect if ν_1, δ_0 increase, a corresponding decrease in ν_2 and μ_2 can feasibly leave the dynamics unchanged. In this example, if we set $\alpha \approx 1.2$ then we arrive approximately at the estimates obtained by EnKF.

Parameter	c	GK	γ_r	μ_2	μ_6	ν_1	ν_2	δ_0
Largest absolute percentage error (%)	1.0	0	0.127	17.13	1.47	20.52	23.91	20.3

Table 9: Largest absolute percentage errors for the different identified parameters in 30 seeded runs for single-measurement cases.

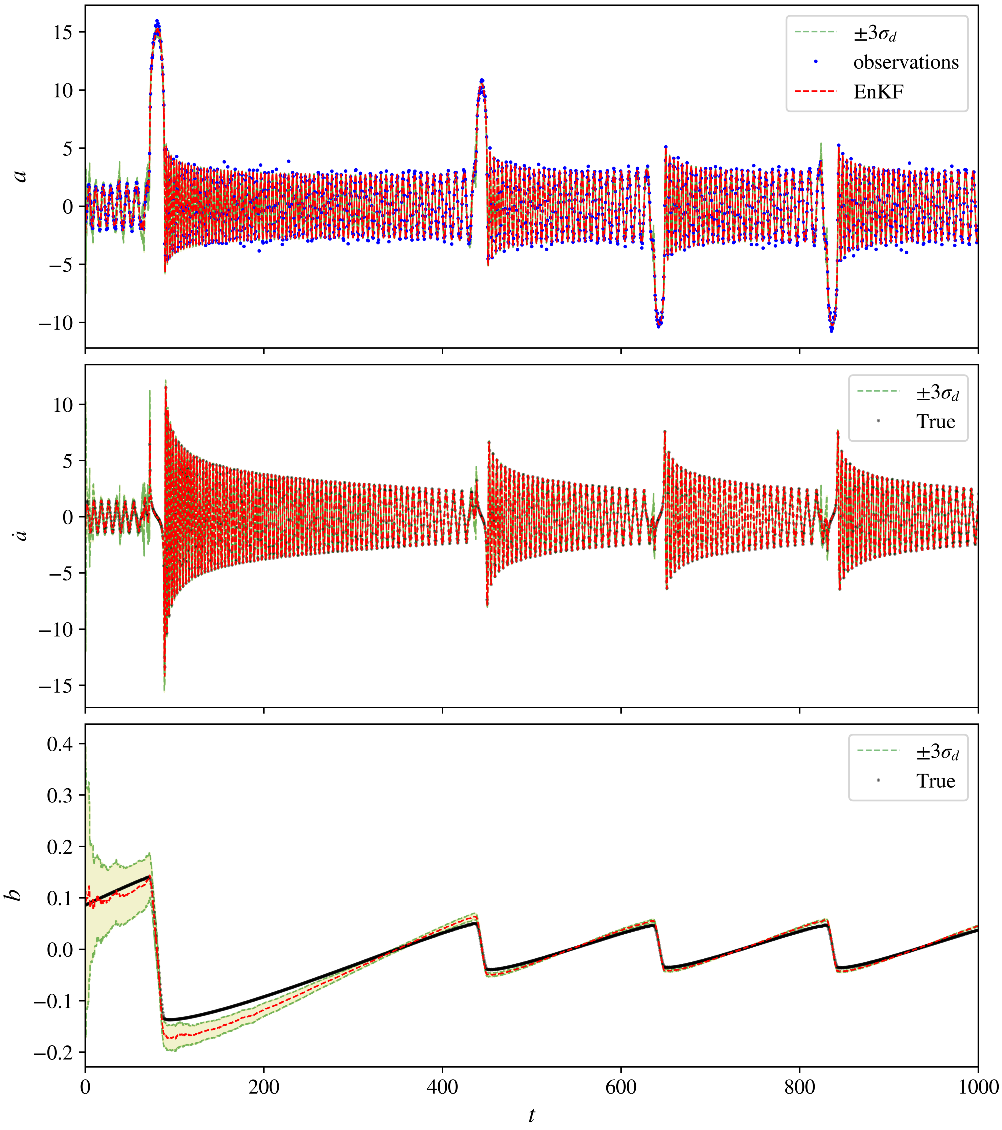


Figure 111: Parameter convergence of the ANAET model with noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The dashed magenta line represents the approximation location of the first crash in amplitude of b.

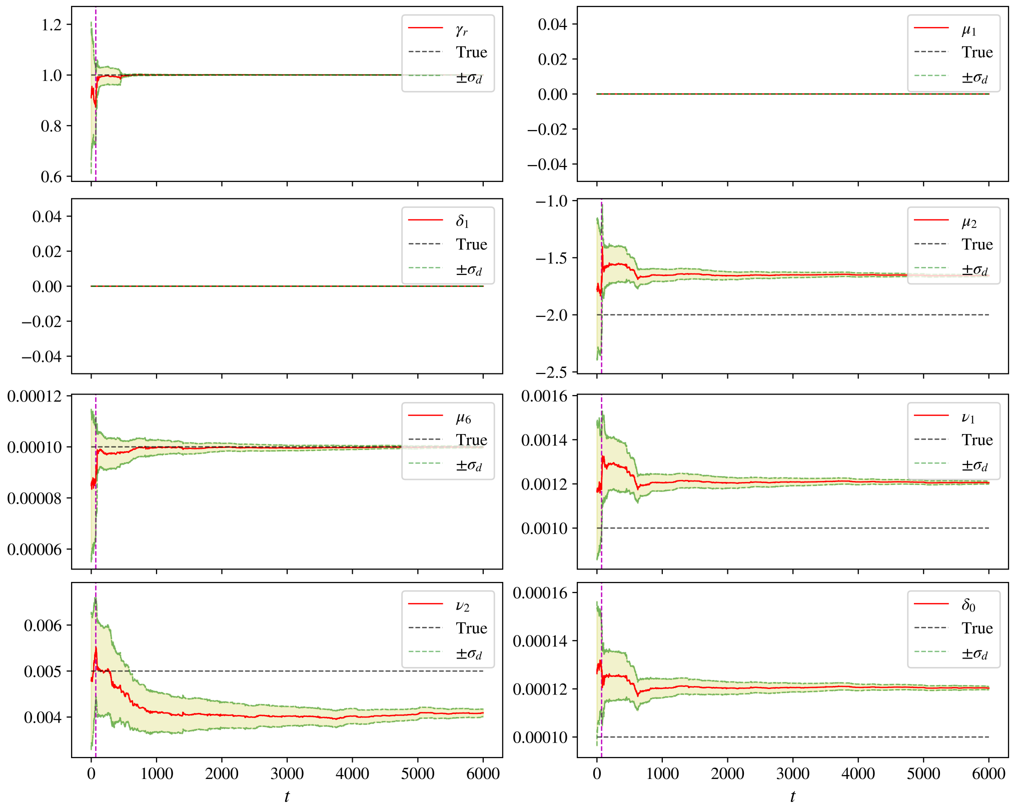


Figure 112: Parameter convergence of the ANAET model with noise $\delta = 0.1\tilde{\sigma}_i$ and sampling rate $\nu_f = 6$. The dashed magenta line represents the approximation location of the first crash in amplitude of b. The parameter convergence here shows a seed which converges to a set of transformed parameters.

7.1.5 Step to experimental measurements

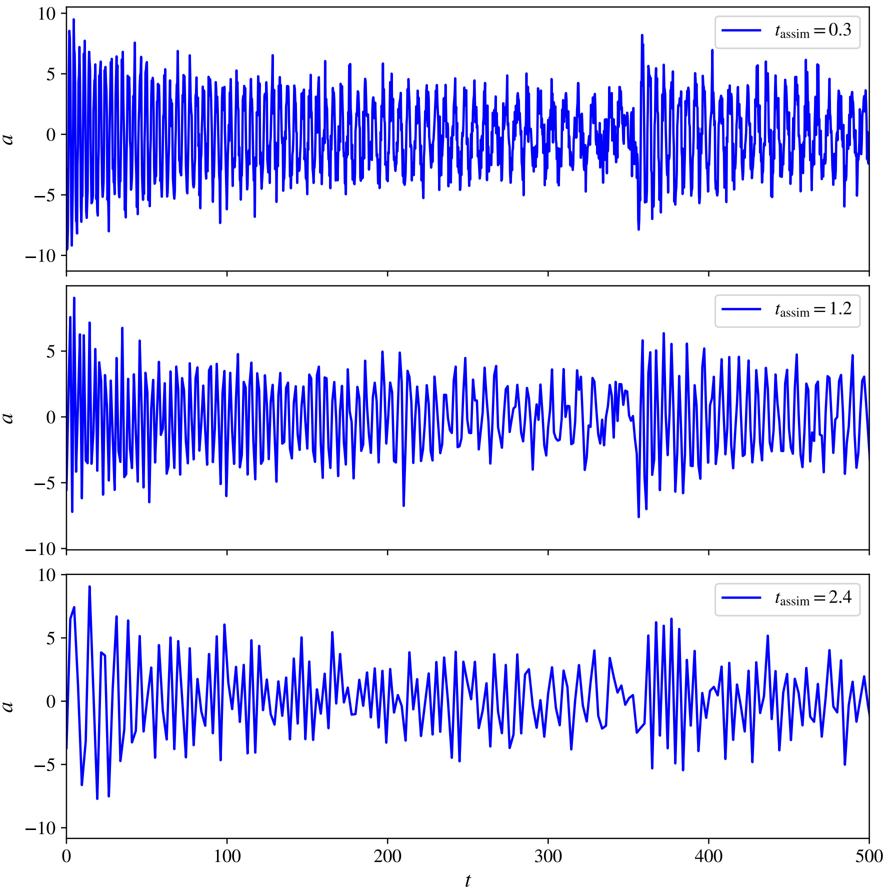


Figure 113: Example of resulting downsampled $\dot{a}$ for different observation rates with additive noise set to 10% of the signal standard deviation.

	a_0	v_0	b_0	γ_r	μ_1	μ_2	μ_6	ν_1	ν_2	δ_0	δ_1
Value	3.983	-5.439	-1.370	1	0	-2	1×10^{-4}	1×10^{-3}	5×10^{-3}	1×10^{-4}	0

Table 10: Second set of initial conditions used with the ENKF method for the ANAET model.

We now attempt to validate the performance of EnKF on observational data-sets which are designed to resemble experimental measurement. In the best case, this will constitute observations corresponding to $\dot{a}$ and b taken from a measurement of the magnetic field and soft X-ray of a sawtooth instability as discussed in the comparison given in § 2.4.3. As we intend to make this relevant to experiment, we will use noisy observations in all cases. While we attempt to mimic experimental conditions as closely as possible, there are some notable caveats when applying to real experimental data. The first of these is the role of process noise in EnKF when there will undoubtedly be non-modelled physical effects which will be discussed in more detail later. The second of these is the correct determination of the type of noise (observational, additive and multiplicative) that is present in experimental signals. In this section we shall assess the convergence of the parameters with different degrees of observational noise and progressively coarser assimilation time-steps to note the limits of the approach.

To increase difficulty of the problem, we assume that we are relatively uncertain of the initial conditions in the parameters allowing the initial guesses to be drawn from Gaussian distributions with standard deviation equal to the magnitude of the true values given in Table 10. The initial uncertainty in each state remains fixed at $[2^2, \delta^2_{\text{obs},v}, \delta^2_{\text{obs},b}]$ where $\delta_{\text{obs},v}$ refers to the observational noise in b and similarly $\delta_{\text{obs},b}$ is the observational noise in a. The initial values of $[a, v, b]$ are set to the first observation for v and b, and $a_0 = 0$. The initial condition and uncertainty in a are chosen as we expect the behaviour of the ANAET model to be a particle in a symmetric potential well so it is reasonable to assume that it will be located around the origin. The initial conditions for the integration of observational data are chosen to start shortly after a crash when the oscillations are fastest. For a fixed assimilation time-step, this results in the fewest samples per oscillation. Examples of a section of observational data are shown in Figure 113. Lower sampling rates can spuriously change the apparent frequency of the trajectory if there are insufficient observations per oscillation.

Again to integrate the ANAET model we use a time-step of $\mathrm{d}t_{\text{phase}} = 0.001$, referring to the time-stepping of the numerical model between assimilation time-steps. When assimilating observations, we perform assimilation at time-steps $t_{\text{assim}} \in [0.3, 0.6, 0.9, 1.2, 1.8, 2.4]$. Uniformly downsampling this way produces a poorer resolution in $\dot{a}$ than b as the time scales in a are faster than b. For varying degrees of noise we use randomly distributed noise with standard deviation given by a fraction of the standard deviation of either $\dot{a}$ or b. We choose the fraction of noise to be within $[0.1, 0.2, 0.3]$ of the respective signal standard deviation. Given that the observational noise varies, we also set $\boldsymbol{R}$ to be the true variance of the observational noise added for each case.

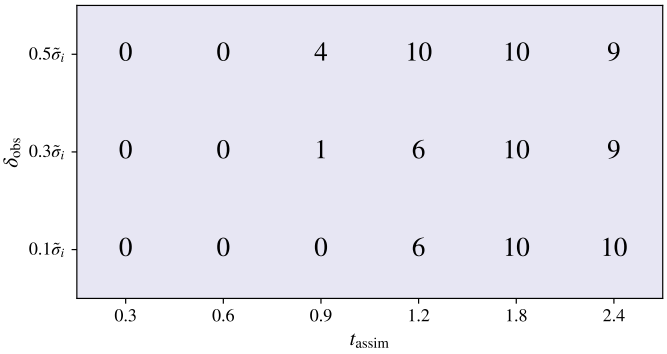
Given that the degree of noise varies throughout benchmarking, assessing the quality of fit based on the agreement to observations is challenging. For this reason we assess only the convergence of the parameters after a fixed time period for varying degrees of noise and sampling rates, assuming that identifying invariant systems is less likely in the double measurement case. To improve the robustness, we also run each of these cases with 30 random seeds which varies the initial guess given to EnKF as well as the initialisation of the ensemble members. During assimilation of the different seeds, several runs diverge, shown in Figure 114. At least in this case the divergence of the filter appears to largely depend on the assimilation time-step and not the degree of noise added. This is not entirely surprising as the initial ensemble may contain parameter values which cause divergence of the filter and cannot be corrected until assimilation. Even if those measurements are noisy they still are able to correct specific parameters to prevent instability. It should be noted that at coarser sampling rates, fewer runs converge meaning the calculation of the means later are over fewer seeds.

Figure 115 shows the calculation of the mean parameter values for the different seeds which do not diverge. The colour scheme is centred around the respective true parameter value with whiter squares representing final parameter values closer to the true value. We appear to have good convergence of most parameters for assimilation time-steps ≤ 0.6 and degrees of noise ≤ 0.3. The main exception to this is ν_2 which generally fails to converge to the true value in the majority of cases. This seems to have little impact in the overall quality of the fit. For higher degrees of noise and coarser sampling rates there is substantially less consistency in the convergence of the parameter, though often the means are skewed by runs which diverge far from the true value. Divergence of the results could be improved by constraining μ_6 corresponding to $\dot{a}a^6$ and this notion will be used in the following sections.

A closer view of the parameter spread is given in Figures 116, 117 and 118 showing the boxplots for each parameter for fixed noise fractions of $0.1\tilde{\sigma}_i, 0.2\tilde{\sigma}_i$ and $0.3\tilde{\sigma}_i$ respectively. In each of these plots, the y-scale has been limited to ±200% of the true parameter value for clarity. These figures highlight the tendency for the mean to be skewed by outliers, even at high sampling rates and low degrees of noise. In many cases the median, indicated by the horizontal red lines, appears to be closer to the true parameter than the mean (green triangles). For all degrees of noise, we can see that oddly $t_{\text{assim}} = 1.8$ has a much higher error than subsequent coarser steps. This is likely an artifact of the either the smaller number of seeds averaged over or the particular set of seeds that diverge in this instance.

For $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ shown in Figure 116 the identification of some parameters appears to remain robust even for sparser sampling rates. However, for sparser sampling rates it is challenging to correctly identify true values for μ_6 and ν_2. This appears to remain the case for higher degrees of noise shown in Figure 118, where below the assimilation time-step of $t_{\text{assim}} = 1.2$ we still have promising parameter convergence. This percentage of noise is well in excess of anything that is likely to be encountered experimentally and provides a promising result for genuine application to experiment.

Finally, Figures 119 and 120 show the same validation in the single-measurement case when either $\dot{a}$ is observable or b. For the case of single measurements for $\dot{a}$ we still have excellent parameter convergence for $\delta_{\text{obs}} = 10\%$ and this remains the case for higher degrees of noise (not shown), though the convergence is unsurprisingly not as good as the two-state measurement case. On the other-hand, even for lower degrees of noise single measurements of b shown in Figure 120 has poor parameter convergence in every parameter. This is hardly surprising as the noise present in the observations obscures the small feedback of a in b. The results are re-assuring on the whole, suggesting that single measurements of the magnetic field alone could be sufficient to fit models, even when that data is noise and sparsely sampled. Unsurprisingly though EnKF does benefit when the observations are of the fastest time-scales in the system. Successful fits of the ANAET model to soft X-ray measurements alone are unlikely to be successful as the high degree of noise pollutes any of the fast-scale coupling of a into b.



δ_{obs} \ t_{assim}	0.3	0.6	0.9	1.2	1.8	2.4
$0.5\tilde{\sigma}_i$	0	0	4	10	10	9
$0.3\tilde{\sigma}_i$	0	0	1	6	10	9
$0.1\tilde{\sigma}_i$	0	0	0	6	10	10

Figure 114: Counts of the number of diverging runs out of 30 seeds for varying sampling rates and noise. The results are displayed on a grid and each number corresponds to the number of diverged run.

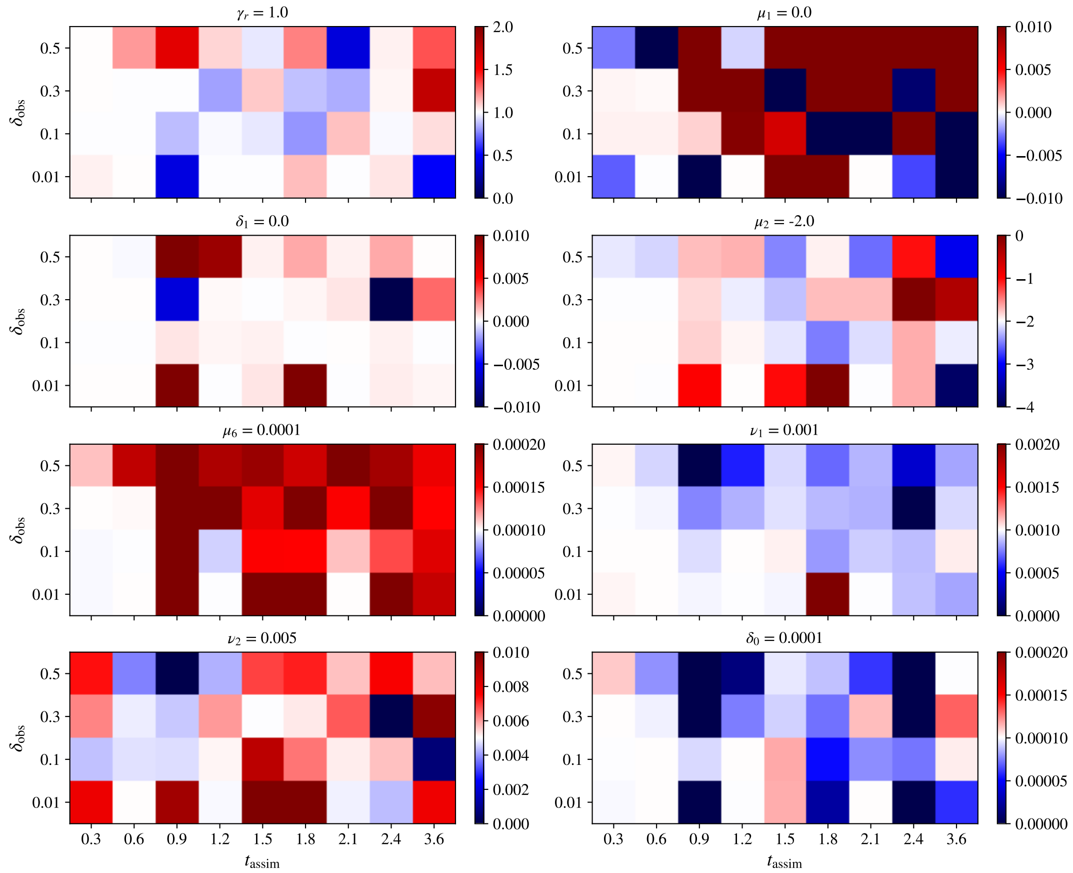


Figure 115: Mean values over all non-divergent seeds for all parameters for varying noise and sampling rates. The colourbars are all set to a range of ±100% of the parameter's true value except for parameters which should be zero are set to have minimum and maximum scales ±0.01. The title shows the parameter's true value, and in each subplot white indicates means close to true parameter value.

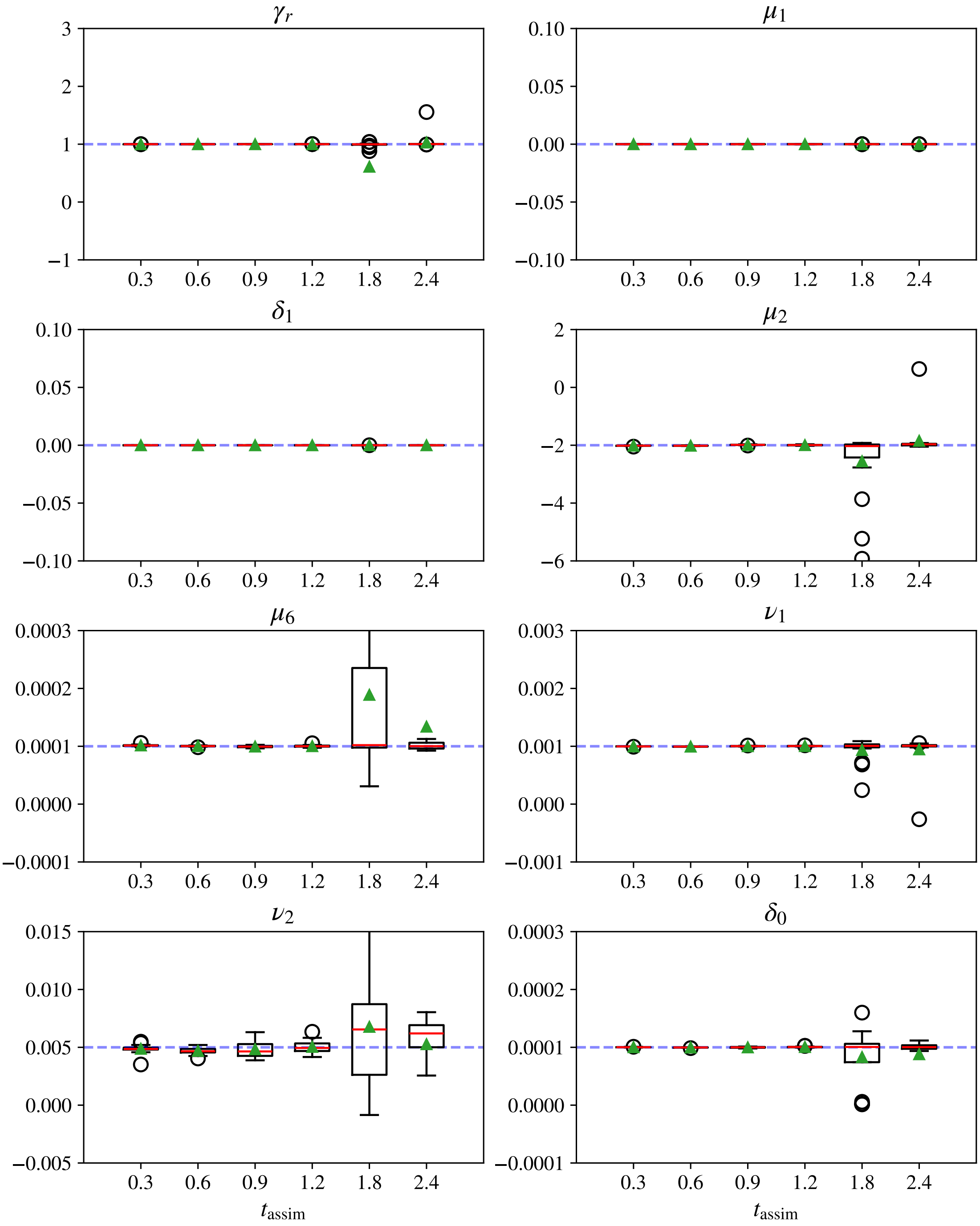


Figure 116: Boxplots for fixed noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ for varying assimilation time-steps. Medians are shown as solid red lines and means as green triangles and the dashed blue line is the true value of the parameter. The results here are for observations of $\dot{a}$ and b.

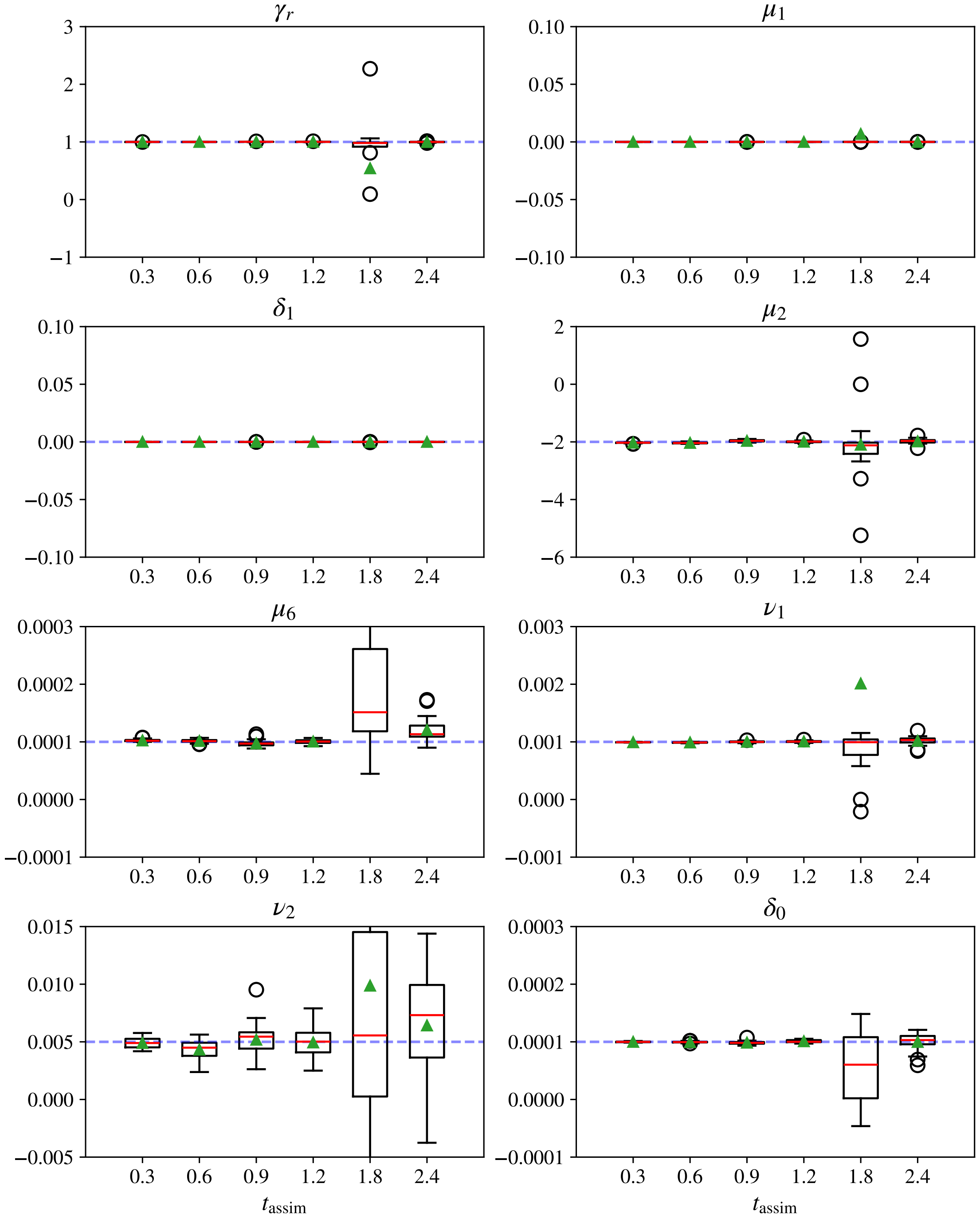


Figure 117: Boxplots for fixed noise $\delta_{\mathrm{obs}} = 0.2\tilde{\sigma}_i$ for varying assimilation time-steps. Medians are shown as solid red lines and means as green triangles. The results here are for observations of $\dot{a}$ and b.

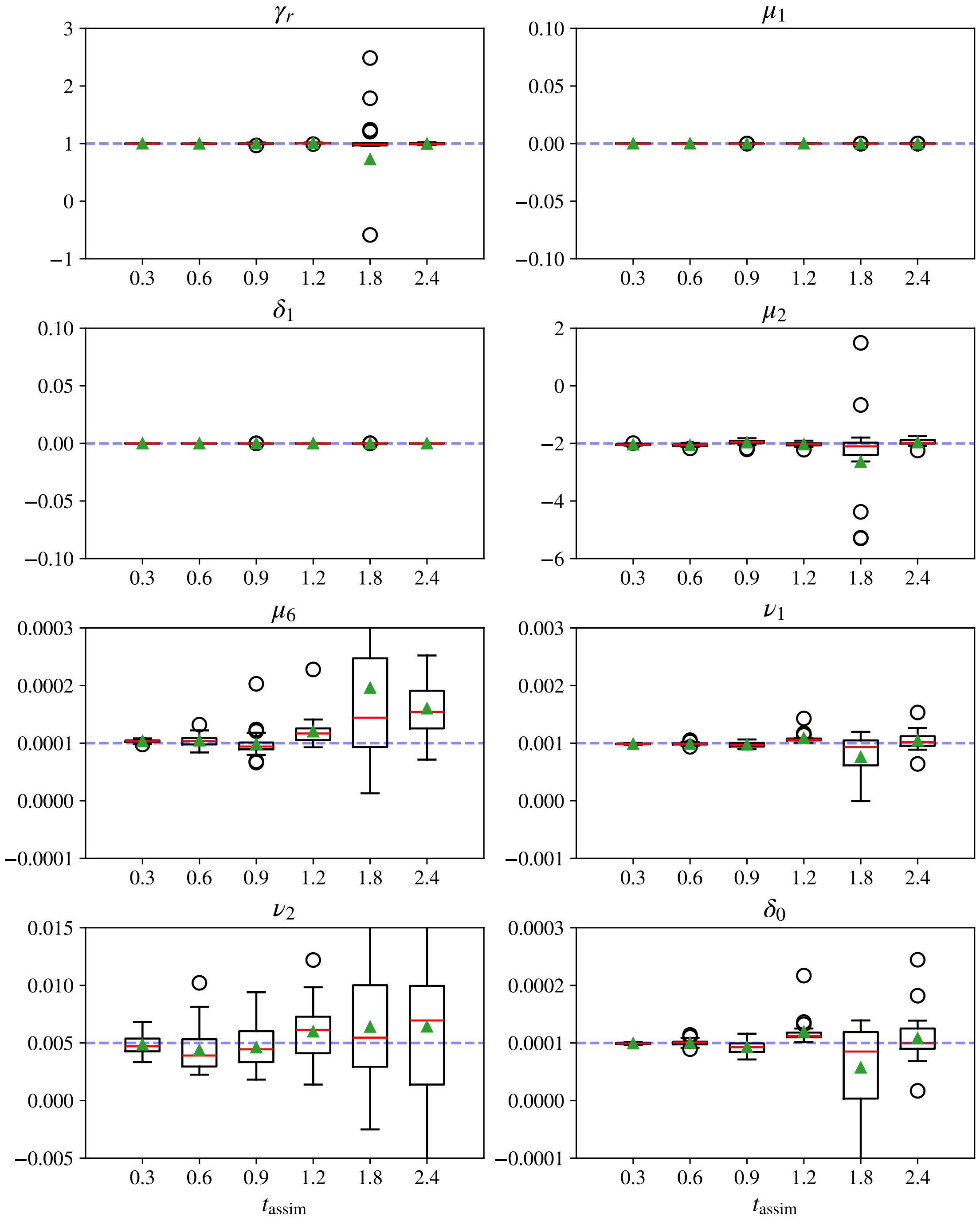


Figure 118: Boxplots for fixed noise $\delta_{\text{obs}} = 0.3\tilde{\sigma}_i$ for varying assimilation time-steps. Medians are shown as solid red lines and means as green triangles. The results here are for observations of $\dot{a}$ and b.

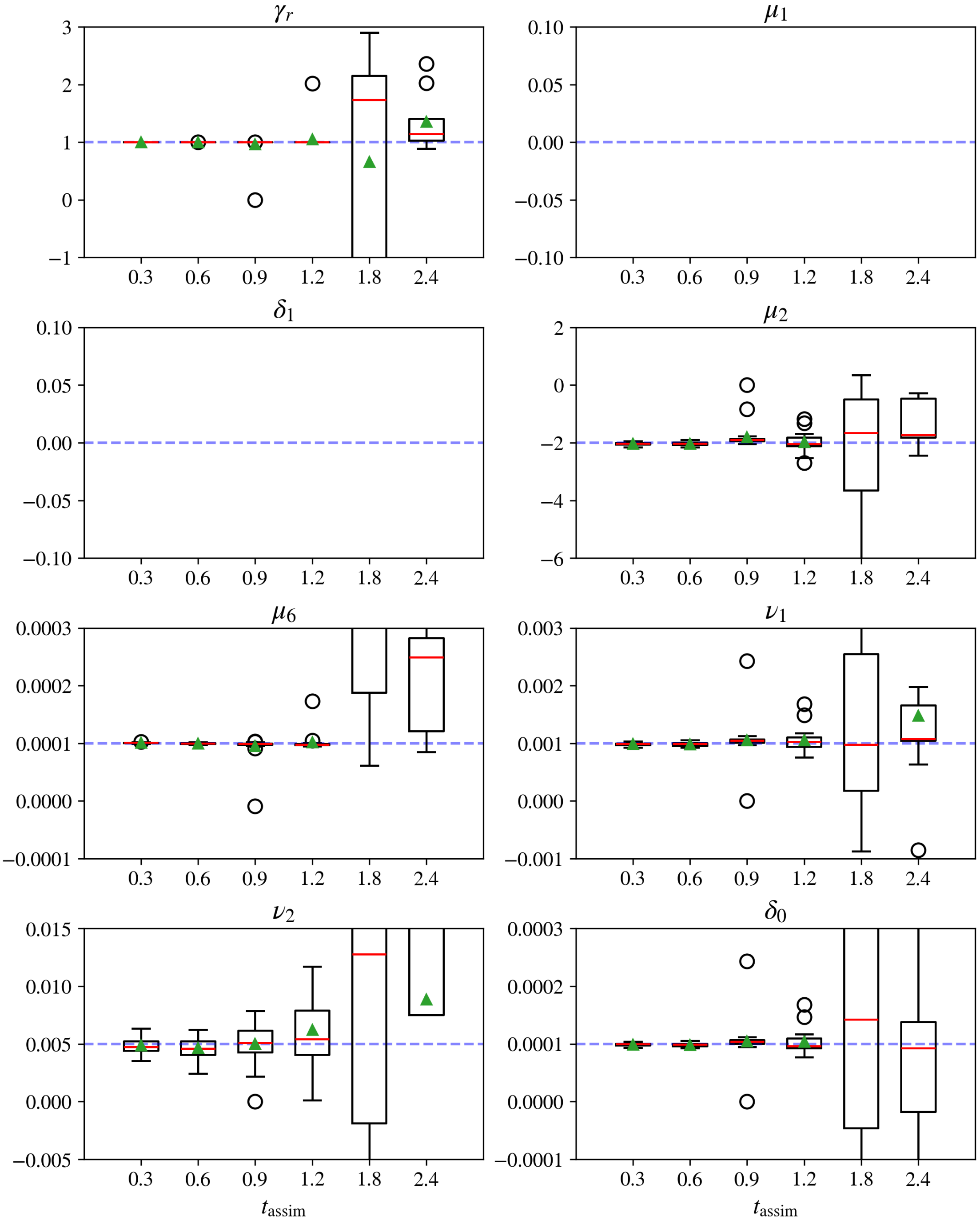


Figure 119: Boxplots for fixed noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ for varying assimilation time-steps with measurements of $\dot{a}$ only. Medians are shown as solid red lines and means as green triangles. The results here are for observations of $\dot{a}$.

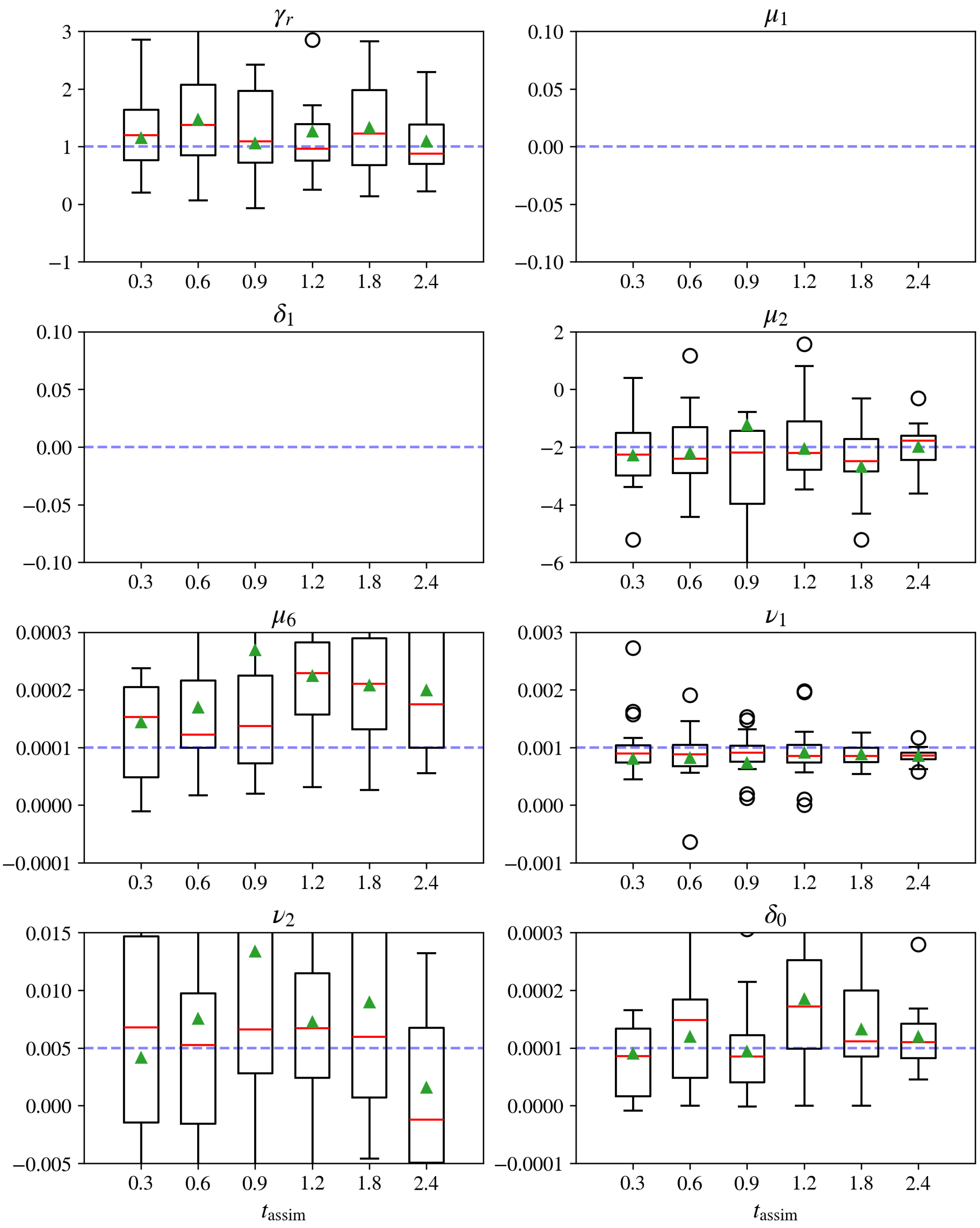


Figure 120: Boxplots for fixed noise $\delta_{\text{obs}} = 0.1\tilde{\sigma}_i$ for varying assimilation time-steps with measurements of b only. Medians are shown as solid red lines and means as green triangles. The results here are for observations of b.

7.2 Stochastic Integration for the ANAET model

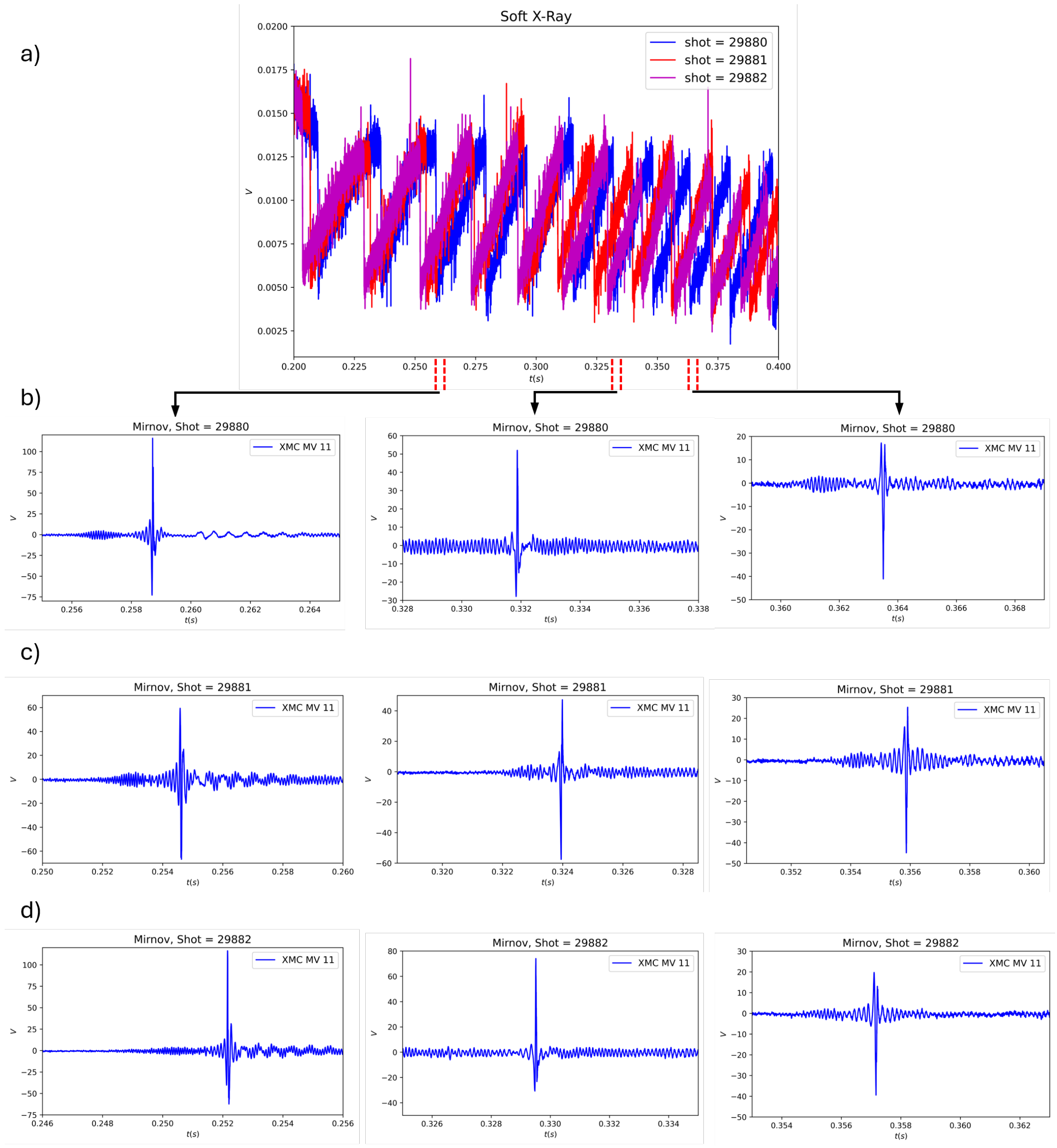


Figure 121: Comparison of magnetic diagnostics taken from the same probe over 3 different similar shots within a MAST campaign. a) shows the soft X-ray measurements with similar sawtoothing events in each shot b) shows the magnetics for shot 29880, c) magnetics for shot 29881, d) magnetics for shot 29882.

The final consideration when fitting the ANAET model to experimental data comes from unrepresented features in the data. Figure 121 shows a comparison of the Mirnov signals and soft X-ray measurements taken from 3 MAST-U shots which are designed to be as close as possible [135]. For each measurement of the soft X-ray measurement from the core, we plot the corresponding Mirnov measurements measured from the same probe. To choose a magnetic diagnostic for this plot, comparisons of all diagnostics are made and we attempt to chose one probe which consistently yields the least noise and most persistent oscillations. Observations of the soft X-rays show that, the period of sawtoothing changes throughout the shot. For EnKF this is important as with fixed parameters, ensemble members will only predict sawteeth on fixed intervals. More importantly, however, is the spiking observed in the Mirnov measurements. For each shot and for different times in the shots, the spiking is similar but can occur at different amplitudes. The ANAET model further will only predict spiking at a fixed amplitude. This leads us to consider introducing model error into assimilation.

Model errors can be incorporated via a discrete stochastic model which evolves every ensemble member within the initial distribution [55][pg 178]. To integrate the ANAET model we use a simple Euler-Maruyama scheme. For the Euler-Maruyama scheme we consider the stochastic differential equation for a system of continuous equations $X \in \mathbb{R}^m$

$$\mathrm{d}X(t) = f(t, X(t))\mathrm{d}t + g(t, X(t))\mathrm{d}W(t) \tag{7.7}$$

where $W \in \mathbb{R}^{n\times m}$ represents a Wiener process implying $W(t+\mathrm{d}t) - W_t \sim \mathcal{N}(\mathbf{0}, \mathrm{d}t)$, $g \in \mathbb{R}^{m\times n}$ is the diffusion coefficient (degree of Brownian motion in the system) and $f : \mathbb{R}^m \to \mathbb{R}^m$ is the deterministic component of our system. The numerical discretisation of this problem over the interval $t \in [t_n, t_{n+1}]$ is given by

$$X_{n+1} = X_n + f(t_n, X(t_n))\Delta t + g(t_n, X(t_n))\Delta W(t_n) \tag{7.8}$$

where $\Delta W(t_{n+1} - t_n)$ and $\Delta t = t_{n+1} - t_n$. The Wiener process essentially boils down to an additive perturbation sampled from a Gaussian distribution of the form $\mathcal{N}(\mathbf{0}, \Delta t)$ where the variance is Δt and the magnitude is given by g. More generalised cases are discussed in [55][pg 178].

In a genuine application of EnKF to experimental data, quantifying the model error is far from clear. The first issue arises in that it is not known beforehand that the chosen ANAC or ANAET models are actually sufficient to describe experimental signals at any parameter values. The second issue is that, as mentioned, errors come from multiple different sources which can be challenging to quantify. For instance, control systems will impact the plasma displacement in a real shot and there is no way, at the time of writing, to quantify this type of error in relation to the ANAC or ANAET models. Further, we have also neglected model error related to the numerical scheme which should arguably be included. Work with the Lorenz model in ref. [29] construct estimates of model error related to integration schemes by evaluating errors between a high accuracy simulation (the truth) and lower accuracy simulations. Again while this would be possible when observations are taken from the numerical model itself, it would not be possible to estimate the model error on actual experimental data as there is no truth. Despite the unknown nature of the model error, in this section we briefly perform some benchmarking of these cases as it is expected that in a real application this will be required.

We now generate observations from stochastic simulations of the ANAET model where we assume only additive noise in the state variables a, v and b. The scale of the noise is chosen to qualitatively alter the period of b to resemble an experimental signal shown in Figure 122. We set $g = [0.05, 0.05, 0.0005]$ as b varies on a much smaller scale than both a and v. Given that there is a stochastic element in our equations which modifies the behaviour, we should expect that we must include some degree of process noise $\boldsymbol{Q}$.

When applying process noise in EnKF there are two ways to apply it

1. Apply it all at once at assimilation time-steps.
2. Apply it on integration time-steps.

For the results described we use the latter approach as it more closely resembles the generation of the observational data.

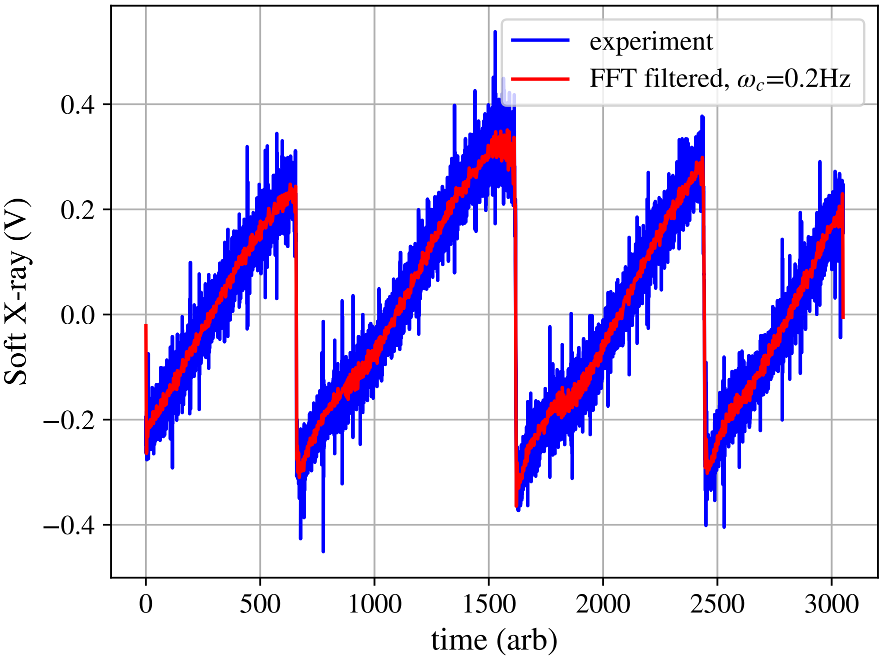


Figure 122: Soft X-ray measurement of the core temperature taken on the MAST-U Tokamak (shot 29880) shown in blue with a FFT filtering shown in red, where high-frequency components have been removed at a cut-off frequency of $\omega_c = 0.2Hz$.

7.2.1 Data generation

	a_0	v_0	b_0	γ_r	μ_1	μ_2	μ_6	ν_1	ν_2	δ_0	δ_1
Value	1	1	0.01	1	0	-2	1×10^{-4}	1×10^{-3}	5×10^{-3}	1×10^{-4}	0

Table 11: Initial condition used for the stochastic simulations. Parameters listed as zero remain fixed at zero.

Observations of sawtoothing signals and Mirnov signals discussed in previous chapters, and shown in Figure 122 highlight some characteristic features of a typical discharge which have not yet been accounted for in the underlying model. These signals do not often display regular sawtoothing occurring at the same amplitude at every crash. In an application to experimental data, this presents challenges for EnKF. For the ANAET model cases considered, parameters are modelled as constants and therefore converge to fixed values. This means that model predictions will always predict sawtooth crashing at constant amplitudes and periods. To generate models which predict non-constant period sawtoothing there are two possible options: additive noise added to the first three state variables or introducing model errors into the parameters. The latter case is much more challenging, as we have unknown parameter values and must make some assumption about the variance increase expected per time-step of these parameters. However, as the parameters can be sensitive to small changes in values this approach is more difficult. For this reason we only consider process noise in the observed variables.

To create the observed data-set, we integrate from the initial conditions shown in Table 11. The additive noise for each variable depends both on the time-step used ($dt_{\mathrm{ANAET}}) = 0.0001$) and the Weiner process coefficient. As b varies on a smaller scale than a and v (this is just a scaling issue and could be changed) we set $g = [0.05, 0.05, 0.0005]$ as this heuristically gives aperiodic behaviour in b. Finally, once the stochastic data has been generated we add observational noise with standard deviation 10% of each respective signals standard deviation. This produces a result shown in Figure 123 which is similar qualitatively to the measurement in Figure 122 from the MAST-U tokamak. Physically, introducing process noise causes aperiodic crashing because random perturbations can cause the particle to be ejected earlier or later from the confining potential well. Comparisons of the attractors created from a Mirnov and soft X-ray measurement from the MAST tokamak to the stochastic integration are shown in Figure 124. The solution behaviour is at least qualitatively similar.

At this point, however, some potential pitfalls of the EnKF approach should be pointed out. As the underlying model is nonlinear, its evolution from an initial Gaussian distribution will almost certainly be non-Gaussian. With process noise, certain ensemble members will crash at different times to other ensemble members. Figure 125 shows an example of $N = 10$ and $N = 100$ ensemble members evolved from the same initial conditions in $\dot{a}_0$ and b but otherwise drawn from a Gaussian distribution. It can be seen that initially the mean of the ensemble provides an accurate estimate of the state, but quickly this breaks when oscillations become out of phase or some ensemble members crash earlier. With EnKF, this is not strictly an issue as the assumption of Gaussian distributions is only relevant during the update step where linear corrections are made based of ensemble spreads [55]. Otherwise the method uses a Monte-Carlo evolution of the ensemble members. However for future prediction without assimilation, Gaussian statistics of the ensemble will not provide accurate uncertainty quantification. On a positive note, as only the update step relies on Gaussian assumptions, joint PDFs could still be constructed from the evolution of the ensemble members for uncertainty quantification as will be discussed later.

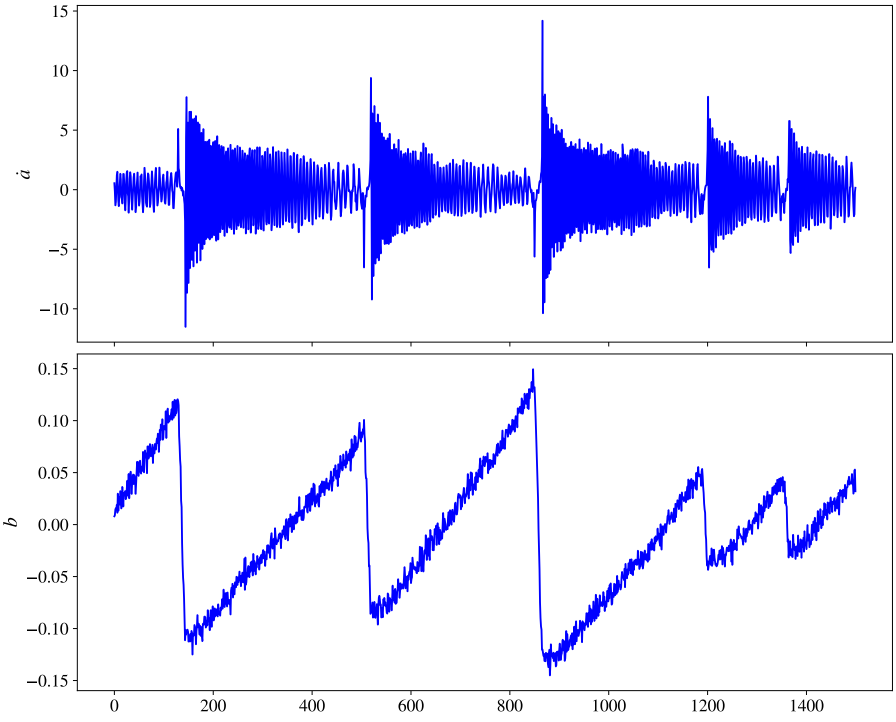


Figure 123: Observable data from the stochastic integration with 10% observational noise added.

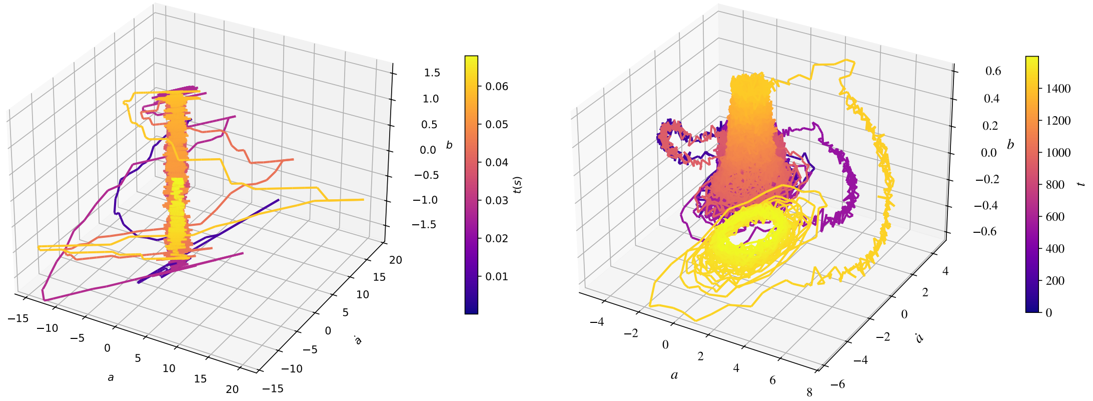


Figure 124: Left shows a plot of a select Mirnov measurement and core temperature measurement taken from MAST (shot 29880), where a is reconstructed from an integral for the Mirnov measurements. Right shows a stochastic integration of the ANAET model with noise added.

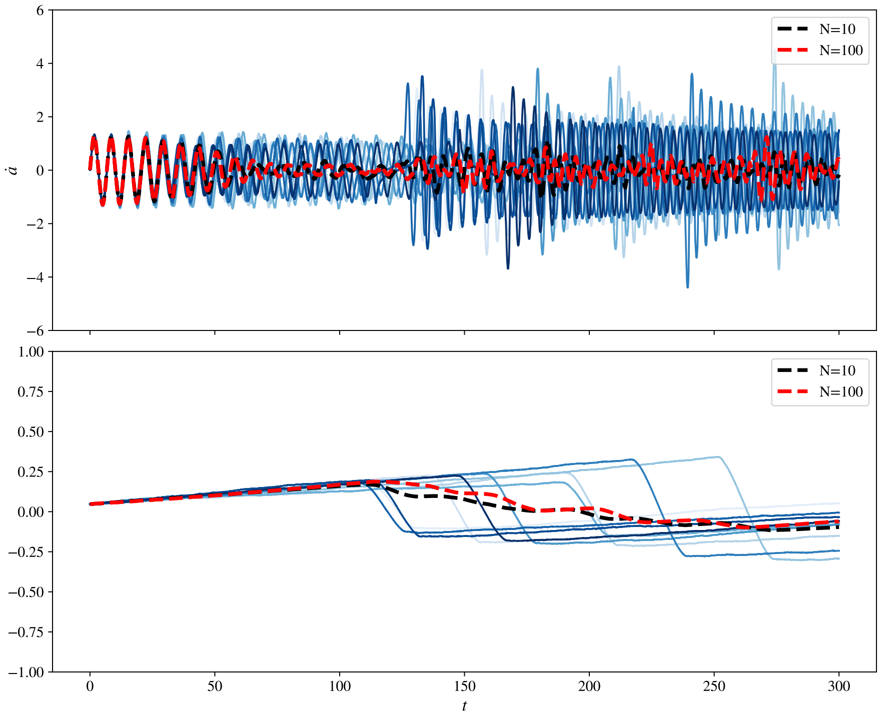


Figure 125: Evolution of different observed ensemble members (10 plotted in blue lines of different shades) with process noise in the state variables for $N = 10$ and $N = 100$ ensemble members. The dashed lines show the ensemble means.

7.2.2 General initialisation of the model

As discussed, sensible initialisation of the priors is an important step in producing physically reasonable ensembles. This is particularly true with the ANAET model where small changes in parameters can cause bifurcations in the underlying model creating non-smooth distributions which have poor Gaussian representations. For a real application to experimental data, we expect to effectively have observations of both $\dot{a}$ and b and we are able to make estimates of the underlying observational error given by the diagnostics benchmarking [166]. For initialisation with non-zero model error we therefore use the following information.

1. The variables $\dot{a}$ and b are observable and these are set to be the first observation at $t = 0$ which is never used during assimilation. The respective uncertainty in $\boldsymbol{P}_0$ of these elements is set to the observational error in each state δ^2_{obs}.
2. a is not known, but we know that the solution is roughly an oscillator centred around $a = 0$, $v = 0$. As such we set the $a_0 = 0$ and let the initial uncertainty by $P_0 = 2^2$ for this variable.
3. We have no knowledge of the parameters, but arbitrary values cannot be used. For sensitivity studies we perturb the true parameter values depending on their respective magnitude. For each parameter we perturb them by a random Gaussian value with standard deviation n times the value of that parameter listed in Table 11. The initial uncertainty is set to reflect this so that the confidence intervals at least bound the correct value.

We will now discuss fits with non-zero model error, initialised as described above.

7.2.3 No process noise

We first begin by attempting to fit EnKF with zero model error to a case where observations are generated with a stochastic integration scheme as described in the previous section. We should expect this to be extremely challenging for EnKF, as the underlying numerical model cannot produce aperiodic crashing behaviour. The convergence of the trajectories shown in Figure 126 showing that while the initial ensemble is spread, EnKF is capable of tracking observations. The parameter convergence shown in Figure 127 shows that parameter convergence is initially close to the true values. However, around $t = 600$ when the ensemble collapses sufficiently, the predictions are over-confident for the numerical model and no longer track the observations. This is then followed by an earlier crash predicted by the ensemble which causes parameters to vary substantially. As there is no error expected in the numerical predictions, the parameters must be varied to match the observations and EnKF is unable to converge to a reasonable solution.

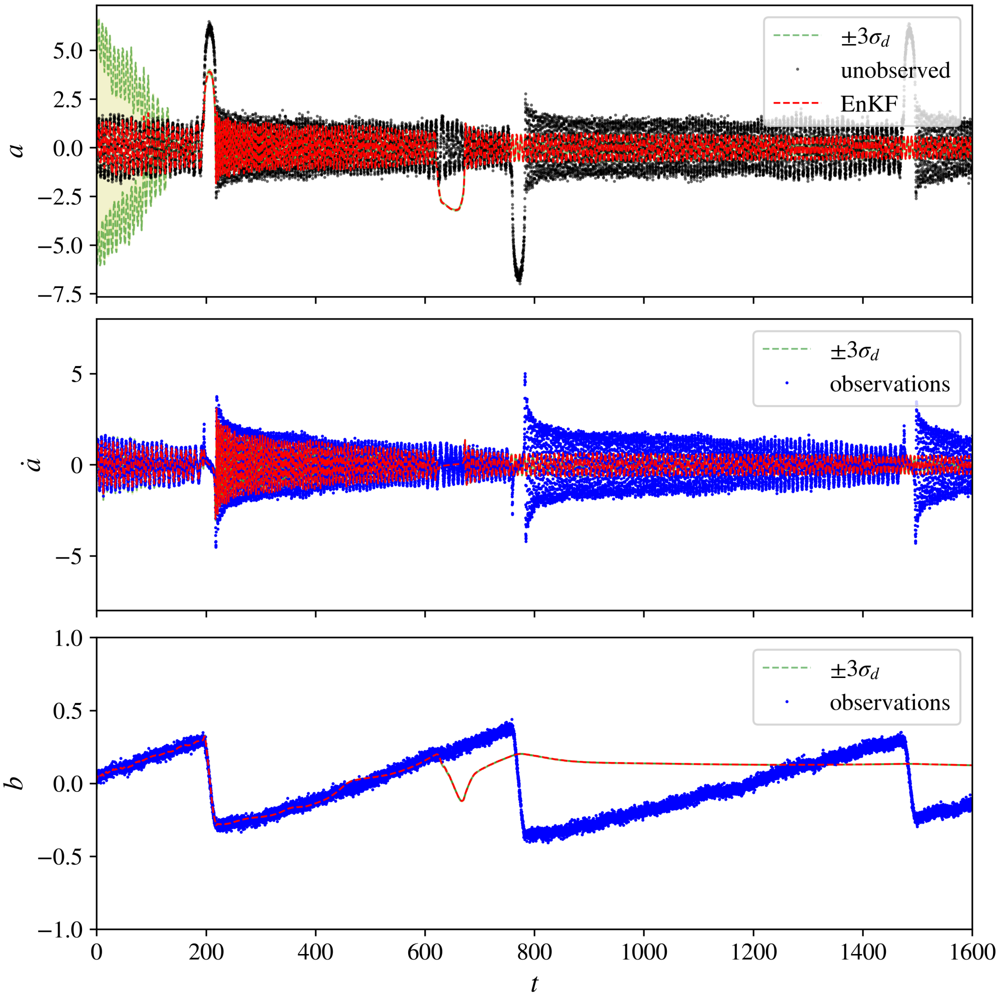


Figure 126: Fit of ANAET model with no model error. The trajectories quickly become overconfident leading to poor fitting of the observational data.

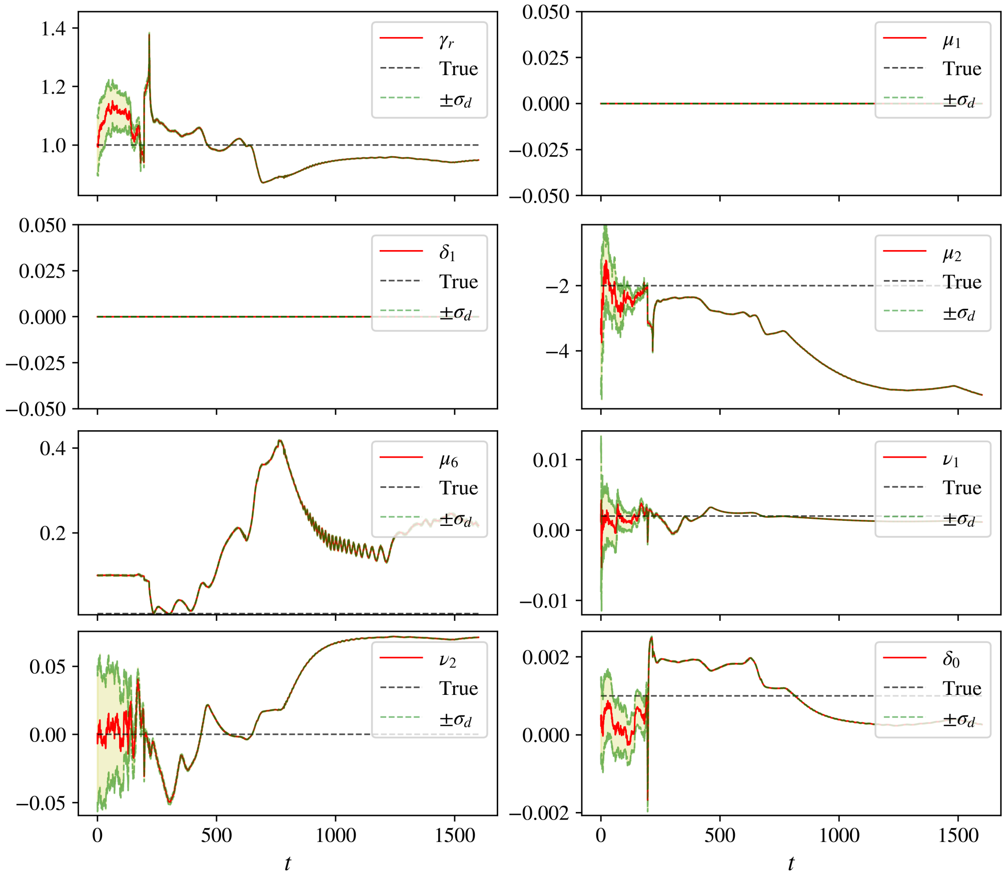


Figure 127: Convergence of ANAET parameters with no model error. Parameter convergence is poor and not close to the true values.

7.2.4 Process noise

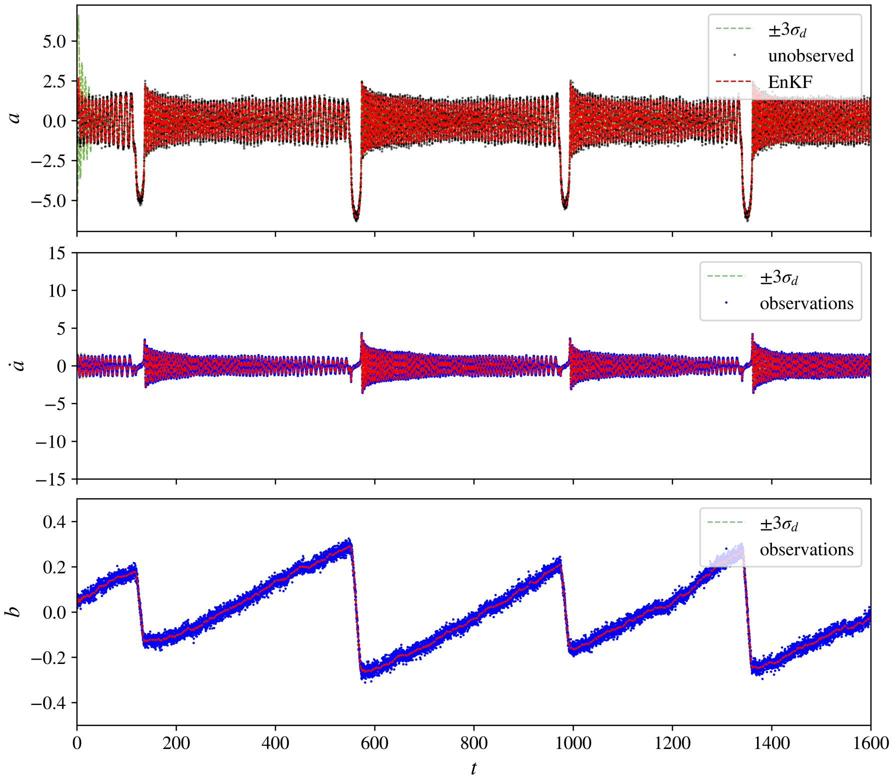


Figure 128: State convergence with expected increase in variance associated to stochastic simulations.

We now look at a case where the initial condition is perturbed by a random Gaussian perturbation with standard deviation equal to the parameter magnitude. We now set the model error elements equal to expected variance increase from the stochastic EM scheme described above. The ensemble members are integrated on a coarser time-step than generated in the EM scheme with $\mathrm{d}t = 0.02$. We therefore set the first three diagonal elements of $\boldsymbol{Q}$ to $[2.25 \times 10^{-6}, 2.25 \times 10^{-6}, 4 \times 10^{-8}]$.

The state convergence is shown in Figure 128 showing that with the described initialisation EnKF is able to successfully track the observations. The inclusion of model error now allows the parameters to be fit as constants, shown in Figure 129. Previously without model error the parameters had to vary so that a varying period could be fit by the observations. We can also see that despite most of the initial guesses being far from their true values, the parameter convergence is still good.

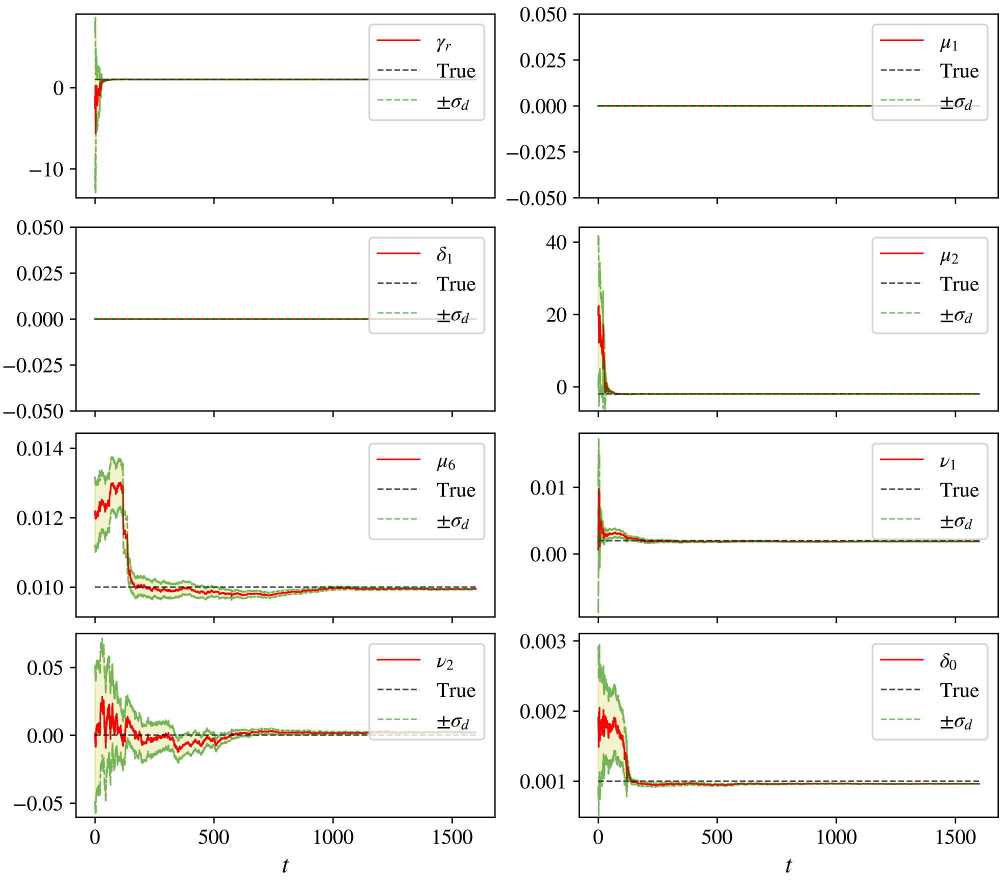


Figure 129: Parameter convergence with stochastic observations when model error is introduced.

7.2.5 Computing uncertainty predictions

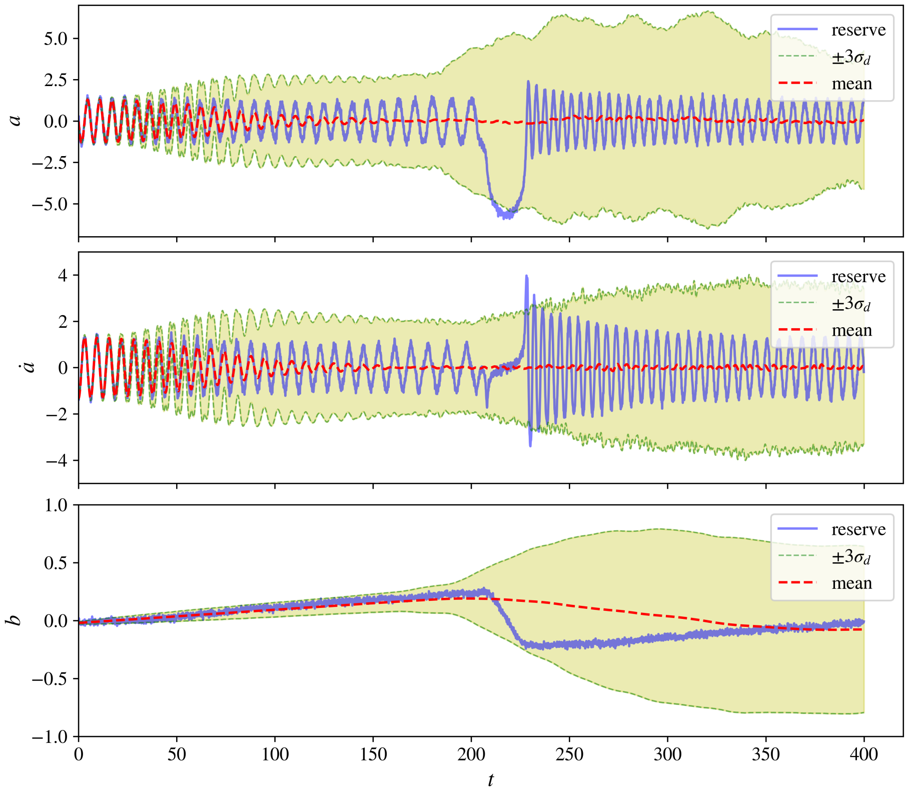


Figure 130: Propagation of ensemble uncertainties without assimilation. The blue data follows from the the end of the training data used to fit the ANAET model. The red line shows the ensemble mean of $N = 300$ members and the green envelope gives the confidence intervals.

As we are now using model error in EnKF, the state predictions will not converge completely and the ensemble spread will increase with a standard deviation $\sqrt{\mathrm{d}t_{\text{phase}}}g$. Following assimilation, we can continue to make predictions with EnKF which amounts to a stochastic integration of the ensemble members. The mean and variance of the ensemble then represent the best estimates. For the previous fit, we continue to integrate the ensemble members for an additional $t = 400$ time units and compare the mean predictions against a reserved section of observations from the stochastic simulation after training. The results are shown in Figure 130 where we can see that reasonable estimates of the mean and uncertainty are only given for very short time-frames. After $t \approx 50$, the ensemble members gradually become out of phase due to non-zero model error and the predictions of many oscillators that are out of phase gives means close to zero. For b we note that the ensemble mean shows a very slow crash between $t = 200$ and $t = 350$ as the ensemble members crash individually implying the mean gives a poor estimate of the ensemble statistics. Also, the variance of b shows ensemble spread significantly higher than $b = 0.5$ which is also artificially caused by the fact that the variance is symmetric around the mean. When some ensemble members crash, the variance increases around the mean and the confidence intervals can span regions of the phase space that ensemble members do not actually reach.

However, as each ensemble member is integrated individually, we can plot the different ensemble members as histograms, shown in Figure 131. The histograms shown correspond to the sampling of all 300 ensemble members at different times. We can see that between $t = 0$ and $t = 200$, many of the ensemble members increase in amplitude. Then between $t = 200$ and $t = 300$ over half of the ensemble members crash to lower values of b and by $t \approx 400$ most members have crashed. Despite the Gaussian assumptions in assimilation we can still gain insight into the uncertainties by looking at the distributions of the ensemble members themselves.

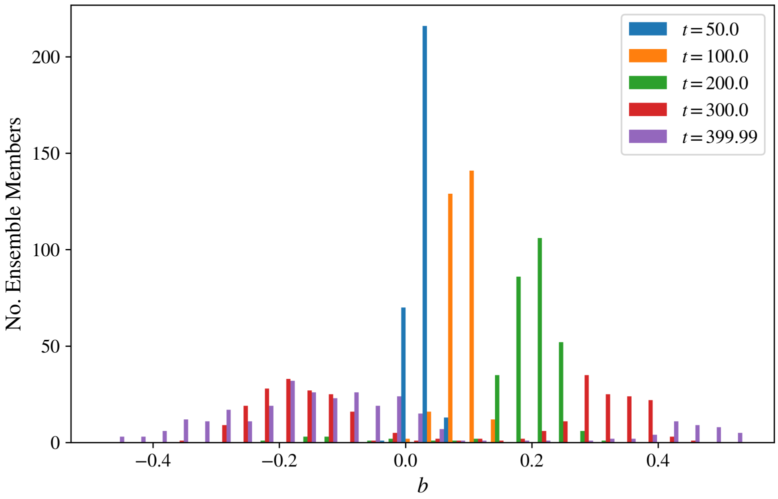


Figure 131: Histograms of the ensemble members taken at different times. As time progressed, more ensemble members crash given by a flattening of the distributions.

7.2.6 Estimation of observational error with FFT

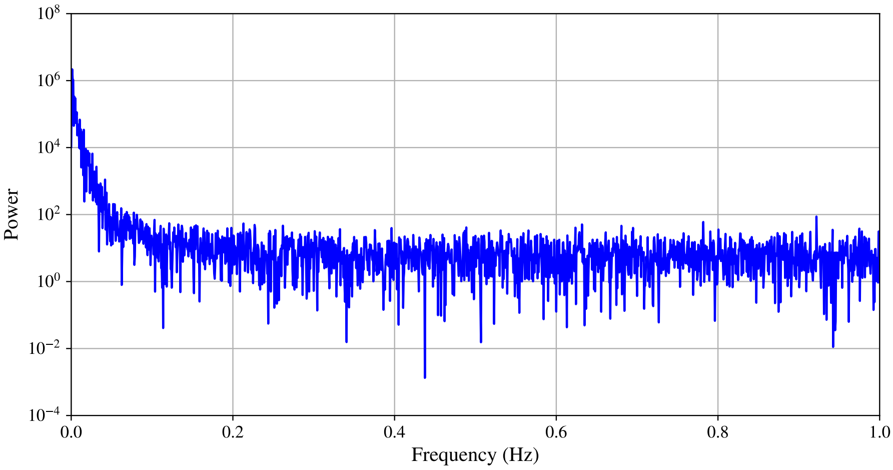


Figure 132: Power-spectral density of b equation after stochastic integration and observational noise added. Only frequencies up to $1Hz$ are plotted as the noise floor continues.

In most applications discussed so far, the observational error $\boldsymbol{R}$ has been assumed known. However, while documentation provides errors for general MAST diagnostics (see ref [166]) individual instrument error can be much higher. It is beneficial to introduce an approach which estimates the observational error for an actual application to experimental data.

If we assume that the noise is high-frequency, then it is possible to make an estimate of the observational noise $\boldsymbol{R}$ through a FFT. We can simply observe the power spectrum of the signal and attenuate frequencies which contribute to the noise floor of the signal. Figure 132 shows the power-spectral density plot for the b equation, highlighting the noise floor for noise estimations. By choosing a cut-off frequency of $0.1Hz$, the standard deviation of the noise can be estimated. Estimates of each degree of noise are shown in Table 12 for each signal. Generally the noise estimates are very close, with a slight over-estimation in noise of $\dot{a}$. This gives a simple and reliable approach for observational noise estimation.

Variable	a	$\dot{a}$	b
Std. Added	0.143	0.103	0.022
Std. Estimate	0.140	0.132	0.021

Table 12: Standard deviations of the random Gaussian noise added to the stochastic integrations of the ANAET model compared to the estimates from an FFT filtering of high-frequency components.

7.2.7 Larger initial uncertainties

Parameter	γ_r	μ_2	μ_6	ν_1	ν_2	δ_0
True	1	-2	0.01	2×10^{-3}	1.5×10^{-4}	1×10^{-3}
IC	1	-15.2	0.1	-0.39	1.02	0.86
Estimate	0.9991	-2.0579	0.0133	2.1229×10^{-3}	-3.4272×10^{-4}	$1.0108 \times^{-3}$

Table 13: Table of the true parameters from which observations are generated, the parameters which are used to initialise EnKF denoted IC and the resulting estimates from EnKF assimilation. μ_1 and δ_1 are not fitted and therefore not listed in the Table.

We now consider a challenging case relevant to experimental conditions with non-zero model error. In a real application of EnKF to experimental data, it is beneficial to use as large uncertainties as possible as initial parameter values are not known. For this test case, we use the same set of observations described previously and set the sampling rate per dominant period to $\nu \approx 60$ which is close to experimental data which has been studied on the MAST-U tokamak with a sampling rate of $\nu \approx 54$ (shots 29880, 29881, 29882). We also add 10% observational noise of each signal standard deviation.

We assume in this instance that our initial guess for the parameters is extremely poor. The initial uncertainty in each parameter is therefore large and set to be equal the standard deviation of the random Gaussian perturbation used for each parameter. The only parameter which is not perturbed is γ_r, as this controls the oscillations between spikes. In a genuine fit to experiment it would be possible to arbitrarily scaling the data in time and amplitude such that $\gamma_r \approx 1$ with an appropriate scaling of the experimental data.

We also implement a lower bound on $\mu_6 \geq 10^{-16}$ such that the term in $-\mu_6 a \dot{a}^6$ is always dissipative. As we implement constraints on this parameter, we also include a small model-error of $\boldsymbol{Q}_{\mu_6} = 10^{-10}$ (the diagonal element of $\boldsymbol{Q}$ corresponding to μ_6) to prevent the ensemble collapsing at the lower bound. This constraint is important, as this term is seventh order and if ensemble members become destabilising EnKF tends to fail to converge. Other unstable choices of the parameters are possible, but ensuring the dissipative term is stabilising at least typically prevents $\dot{a}$ from becoming unstable.

The results of the parameter convergence are shown in Figure 133, with explicit comparison of the initial conditions and final estimates shown in Table 13. We can see that despite poor initial guesses in many of the parameters, we still convergence to very good estimates of the parameters. The main discrepancy is with ν_2 which finds a parameter with the wrong sign. However, integrations of the resulting model still show sawtoothing behaviour on the correct period.

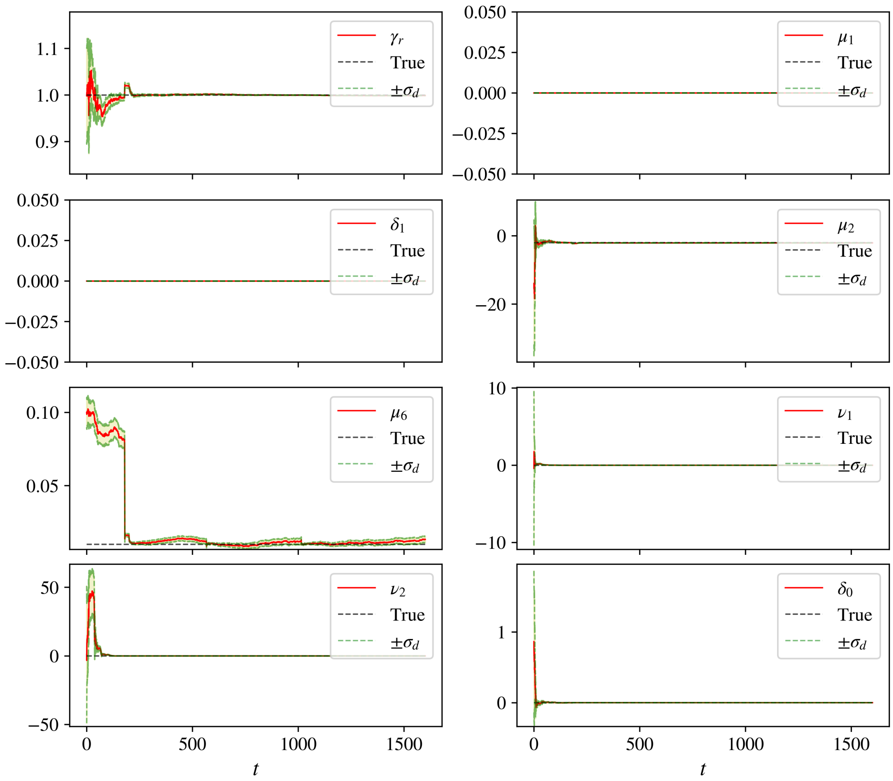


Figure 133: Parameter convergence in the stochastic case where the initial parameter guesses are extremely poor.

7.3 Concluding Remarks

In this chapter we extended EnKF to the more complicated ANAET model in the full-state, double and single measurement cases. The ANAET model represents a significantly more complicated test case as the behaviour is both multi-scale and there are more free parameters to estimate. The ANAET model represents a realistic test-bed for genuine experimental signals with comparisons between experimental signals and stochastic integrations of the ANAET model show qualitative similarity. EnKF shows promising results for estimating the parameters in the full-state, double and single state measurement cases provided that sensible initialisation is provided. Parameter convergence is still good when comparing to scenarios where the signal is very poorly sampled and very noisy. Single measurement cases must be restricted to measurements of state variables in $\dot{a}$, as measurements of b alone do not provide enough information to fix $\dot{a}$.

Extensions to include stochastic integrations of the ANAET model, while preliminary, are an important step in comparison to real experiment. As noted, real experiments often show aperiodic signals and will require the inclusion of model error. A case of attempting to fit EnKF with no model error to stochastic observations showed the method failed to converge and this would likely occur with experiment. We showed that when the variance increase is estimated correctly, EnKF is able to make very good estimates of the parameters even with quite poor initial guesses of some of the parameters. This is also in a case where the sampling rates are equivalent to an observed MAST-U shot of the magnetic field and central core temperature. Also a simple approach using a FFT is outlined to make estimates of the observational noise which will be beneficial on real signals. Finally, we also gave a discussion of uncertainty quantification within EnKF. For nonlinear models, the assumption of Gaussian ensembles is quickly violated without assimilation. Despite this the distribution of the ensemble members themselves still provide information on the number of ensemble members that have crashed. This allows us to construct a probabilistic interpretation of when a sawtooth crash may next appear.

One of the biggest conceptual challenges with EnKF is the choice of initial parameters which is largely unknown. While the optimisation work described by [106] was reproduced, extending this to the ANAET case was unclear for a number of reasons. The first is that, no discussion is given of the ensemble uncertainty when attempting to optimise over the first J observations. Many of the parameters vary on different scales and sensible initial uncertainties must also be specified (whereas with the ANAC model using $\boldsymbol{P}_0 = \mathbb{I}$ is likely fine). The second issue is that the ANAET model is multi-scale and many of the parameters associated to the slow-mode do not converge until approximately one slow-period of the system is assimilated. This would make the optimisation procedure substantially more expensive, as in [106] optimisation only occurs over the first approximately $J = 12$ observations whereas a slow oscillation will easily have hundreds of samples. Alternatively scaling the data could be explored so that certain parameters, such as γ_r, would be approximately 1. This would reduce the complexity of having large uncertainties in all the parameters. Aside from this it seems that in many cases we can still use large initial uncertainties and still see successful convergence of EnKF. Other approaches such as particle filters implemented in Julia by ref [162] have also performed assimilation with the ANAET model, and could be viable alternatives when parameters are poorly known. This is because particle filters do not assume Gaussian distributions and perform a re-weighting process for each particle (ensemble member). This could be be beneficial, as ensemble members drawn from non-physical priors can be assigned a zero weight. While some attempts were made at using particle filters, time constraints prevented any substantial comparisons being made.

The next logical step is to use EnKF with experimental data, but estimation of model error remains a significant challenge. This is because both the sufficiency of the ANAET model is not known, and the accuracy of the model depends on the parameter estimates themselves. Without any knowledge of the parameters beforehand, it is difficult to anticipate the model error. While extensive attempts have been made to find good data-sets to test on, finding suitable shot candidates from MAST/MAST-U is very difficult and likely requires input from an expert familiar with the diagnostics. While sawtoothing events may be present in shots, there can often be large sections of poorly resolved Mirnov signals with only one or two sawteeth apparent. Figure 134 shows a comparison of the magnetic diagnostics from the same location in the MAST tokamak with the corresponding soft X-ray measurements for 3 different shots in a MAST campaign which were designed to be as similar as possible [135]. Despite the fact that all shots show a sawtoothing event, there are many cases where the magnetic signal is effectively just noise preceding a sawtooth crash, such as shot 29881. Even for shots where the signal is relatively well resolved as in 29880, this only persists for a short section of the signal shown in Figure 135. Aside from this, the oscillation amplitude between sawtooth crashes, shown in Figure 135 can change throughout the shot which makes the design of diffusive terms which are negligible for $|a| < 1$ challenging. Further the spiking observed can also occur at different amplitudes and some allowance for this would have to be made in the underlying model.

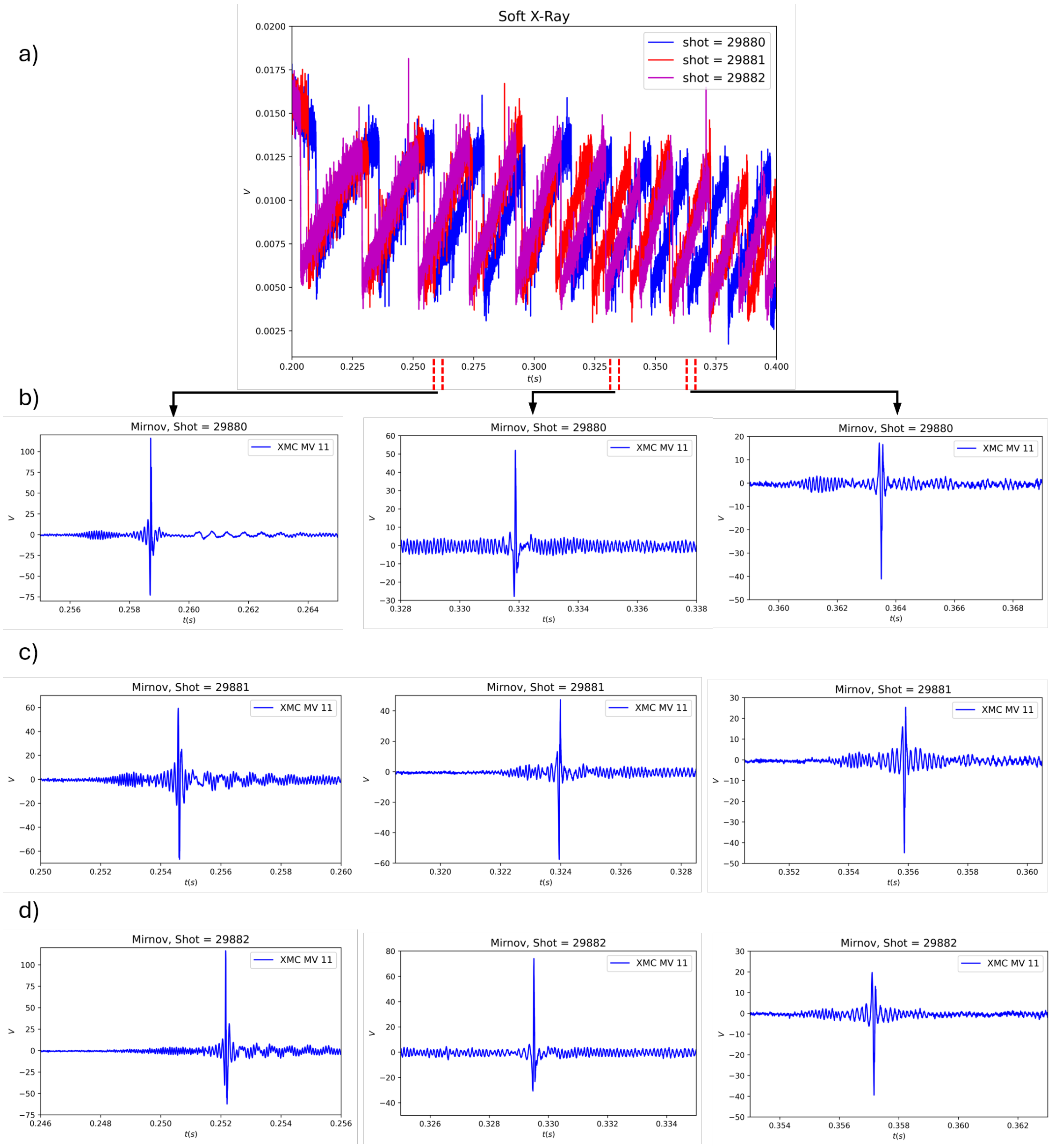
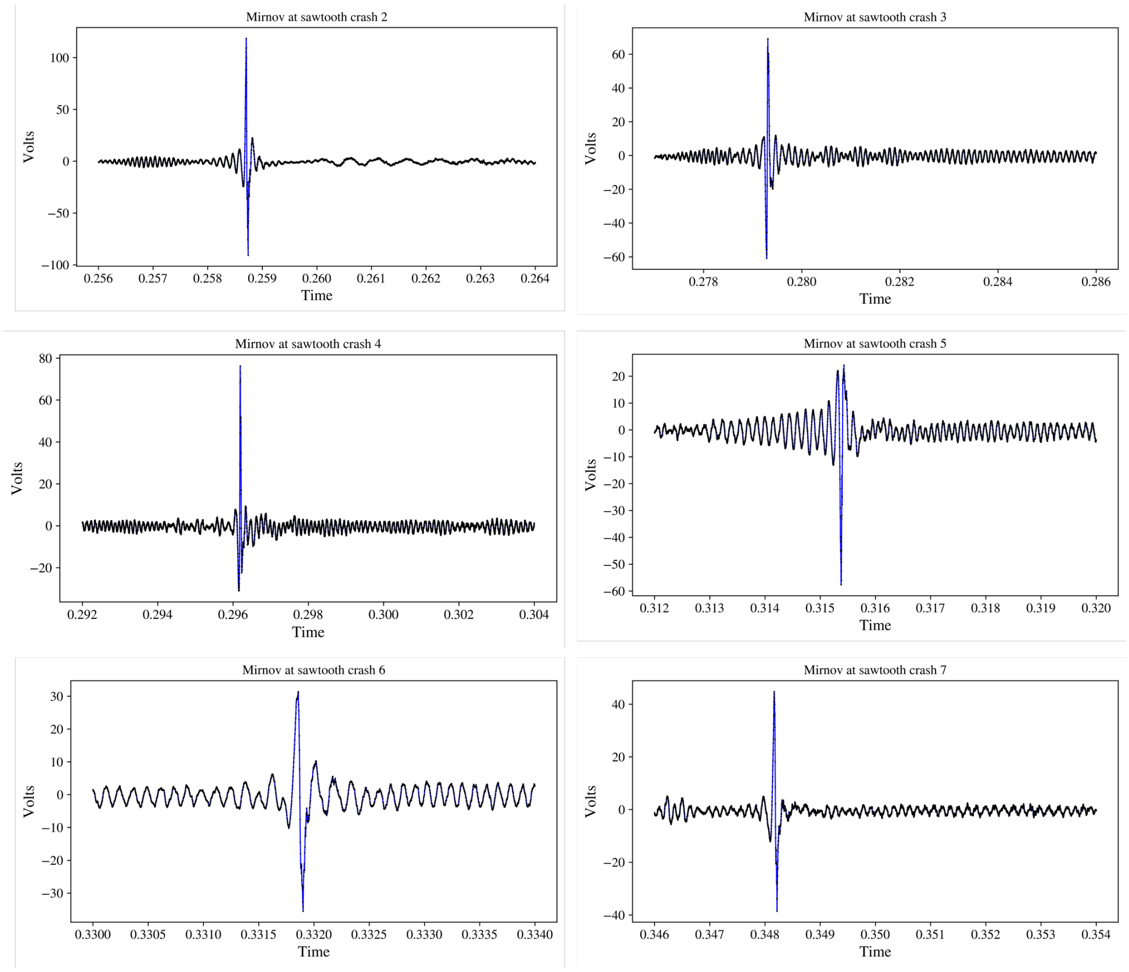


Figure 134: Comparison of magnetic diagnostics taken from the same probe over 3 different similar shots within a MAST campaign. a) shows the soft X-ray measurements with similar sawtoothing events in each shot b) shows the magnetics for shot 29880, c) magnetics for shot 29881, d) magnetics for shot 29882.


Mirnov at sawtooth crash 2

Mirnov at sawtooth crash 3

Mirnov at sawtooth crash 4

Mirnov at sawtooth crash 5

Mirnov at sawtooth crash 6

Mirnov at sawtooth crash 7

Figure 135: Comparison of magnetic diagnostics taken from MAST-U shot 29880 at different crashes. Each crash number is labeled in the subplot title.

8 SINDy and the ANAC model

In the previous chapters we discussed applications of EnKF for fitting the ANAC and ANAET models to experimental data. While this approach shows remarkable robustness to noise, there are some drawbacks. The first of these is that we typically have no knowledge of an appropriate set of parameters to initialise EnKF. This is important because large model uncertainties tend to create unphysical ensembles, negatively impacting convergence. The second issue is that the ANAC and ANAET models are derived based on general symmetry constraints and observations of tokamak behaviour [72]. As such, the inclusion of other terms may be relevant. For instance, diffusion is introduced to represent non-ideal effects in the ANAET model, but the choice of the diffusion term only requires that the symmetry constraints are satisfied, and it is negligible for small amplitudes.

In this section we briefly explore SINDy applied to noisy, sparsely sampled observations which are generated from the ANAC model. SINDy has some potential benefits over EnKF, particularly that we do not require any prior knowledge of the parameters of the model. Further, we can assume a general function library which obeys the constraints outlined in [72] and therefore does not need to be as restrictive in the selection of terms appearing in the equations. The main limitation of SINDy in comparison to EnKF is the requirement of a complete set of measurements.

As we aim to build a comparison of EnKF with SINDy, we will predominantly be interested in constructing simplified models of the dynamics using SINDy at sampling rates comparable to the observational data-set used by [115] and those in the previous section with EnKF. For most purposes, this will be about 12 samples per period, though typically this is quite restrictive given modern diagnostics on MAST-U [166]. We will then aim to draw comparisons in terms of parameter estimation and state prediction between the Kalman software and SINDy. It is expected that the performance of the Kalman software will be superior because SINDy must estimate a larger number of unknowns in a correlated feature library making for a harder problem to solve. Further, the addition of noise complicates the fitting process for regular SINDy as we must obtain derivatives. However, we will find that with a sound choice of optimiser and sampling method, we can obtain good results from SINDy for arguably harder problems in high-noise regimes.

We therefore aim to answer the follows questions in this section:

1. How does the performance of EnKF compare to SINDy for parameter inversion with the ANAC model?
2. How robust is SINDy to noisy signals from the ANAC model?
3. How robust is SINDy to poorly sampled signals from the ANAC model?

8.1 Conventional SINDy

8.1.1 SINDy with noiseless observational data

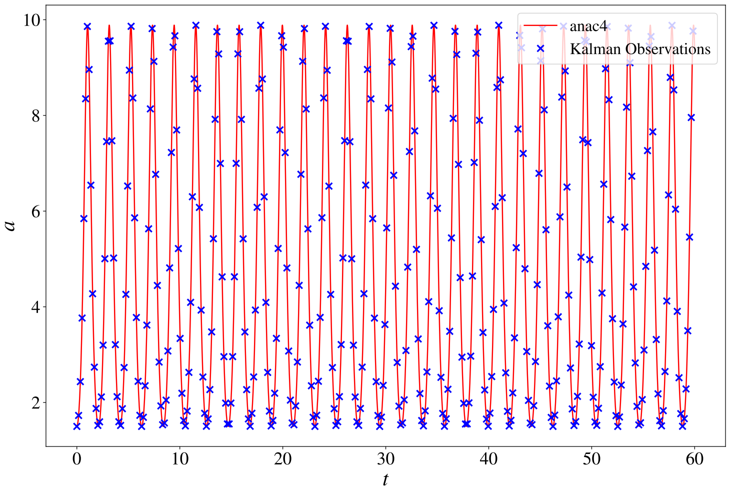


Figure 136: Integration of the ANAC model shown in red with observations taken at $\nu \approx 12$ shown in blue crosses.

We start with an artificial noiseless observational dataset which is generated from the integration of the ANAC model

$$\ddot{a} = \mu a + 2\sigma a^3, \tag{8.1}$$

at $\mathrm{d}t = 0.005$ with $\sigma = -0.1$ and $\mu = 10$ and we introduce the variable $\dot{a} = v$. The data is then down-sampled such that there are approximately 12 samples per period and $\mathrm{d}t_{\mathrm{assim}} \approx 0.16$. In the initial example, we consider only noiseless observational datasets, shown in Figure 136. The first challenge then is to obtain accurate model identification with SINDy for these sampling rates. As can be seen, the observations are sparsely sampled in time and thus result in poor approximations of the derivatives using finite differences. We display all results in terms of the dominant period of the data which we label Λ and from a FFT is set to $\Lambda \approx 2$.

We start understanding the initial limitations of SINDy by performing parameter sweeps with the sequential thresholded least squares optimiser. From previous results in §4 we should expect that the weak form will be required. To evaluate model performance we consider two different performance metrics. The first is the mean-squared error

$$\mathrm{MSQE} = \frac{1}{nm} \sum_{i=0}^{i=n-1} \sum_{j=0}^{j=m-1} (\dot{X}_{ji} - \hat{\dot{X}}_{ji})^2, \tag{8.2}$$

where $\dot{X}$ are the true derivatives and $\hat{\dot{X}}$ the predicted derivatives from SINDy. The result is averaged over all n equations. For the ANAC model, this constitutes $n = 2$ when written as a first-order system of equations. We calculate the mean-squared error for both the training data and a reserved set of validation data, though we calculate these on noiseless data. It is also common to calculate a mean absolute coefficient error on the non-zero coefficients in the identified model [152]

$$\mathrm{MCE} = \frac{1}{nk_{nonzero}} \sum_{j=0}^{n-1} \sum_{i=0}^{i=k_{nonzero}-1} \frac{|C_i - \hat{C}_i|}{|C_i|} \tag{8.3}$$

where C_i is the array of true coefficients. The result is again averaged over all equations. Note that as SINDy consists of a feature library which contains terms which do not appear in the ANAC model, inclusion of these terms will negatively impact this score.

8.1.2 Identification with STLSQ

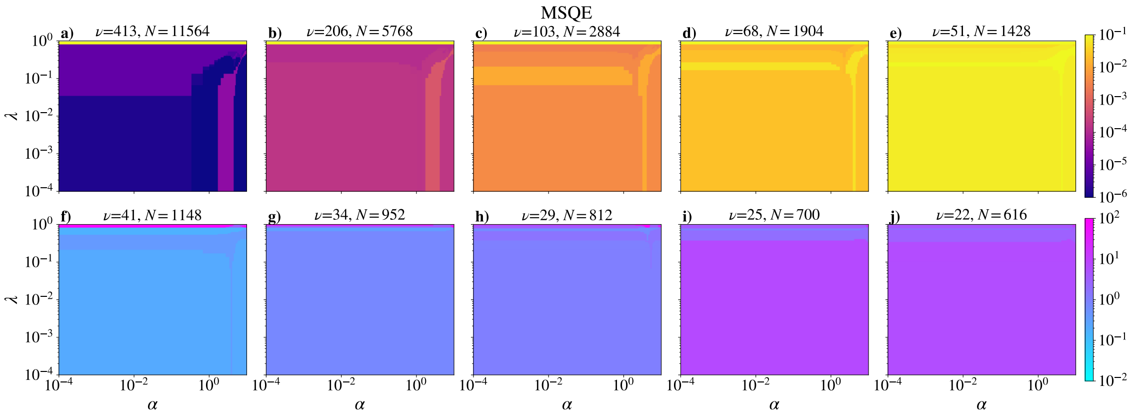


Figure 137: MSQE of the predicted SINDy model on the noiseless training set while varying the coefficient threshold λ and L2 regularisation α. Each box represents the errors at different sampling rates with the total number of samples denoted by N in the subplot title.

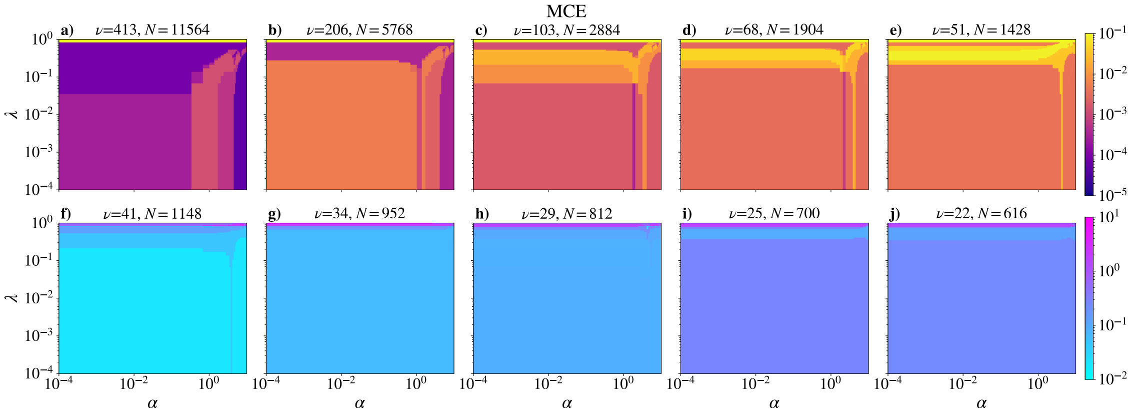


Figure 138: MCE of the predicted SINDy model on the noiseless training set while varying the coefficient threshold λ and L2 regularisation α. Each box represents the errors at different sampling rates with the total number of samples denoted by N in the subplot title.

We first assess the performance of the baseline version of SINDy on progressively coarser datasets to understand the limitations in an optimal case. We begin by generating a dataset from the initial conditions $[a_0, v_0] = [1.5, 0.1]$ with $\sigma = -0.1$ and $\mu = 10$ which results in generically spiky behaviour discussed in ref. [179]. For each sampling rate, we vary the learning parameters corresponding to the hard threshold, λ and the L2 regularisation, α, calculating the aforementioned metrics on the validation set. α and λ are both varied logarithmically between 10^{-4} and 1 in a total of 50 steps. For every SINDy fit we limit ourselves to a third order polynomial library which is a reasonable choice given description of the construction of these models [72]. Higher-order terms are only included when either the potential is expanded to a higher order or diffusive terms are included [179].

Figure 137 shows the calculated MSQE when the sampling rates and learning parameters are varied. The best performance is obviously obtained on the dataset with the highest sampling rate. We see a reduced performance for lower sampling rates, eventually obtaining extremely poor recovery well before reaching the desired sampling rate of $\nu = 12$. From Figure 138 we can see that the coefficient error increases continuously with decreasing sampling rate.

To assess the source of the error, we perform the same parameter sweep by first pre-calculating the derivatives on the data where successful recovery is possible ($\nu = 413$). The pre-calculated derivatives are then supplied to SINDy to avoid calculation of the derivatives on the sparsely sampled data. Figures 139 and 140 show the same parameter sweeps with the supplied derivatives. It can clearly be seen that correct model identification is maintained if high quality derivatives are given. This suggests that the sampling rate is sufficient to resolve the dynamics and an improvement must be made on either the approach or the approximation of the derivatives. In fact, very few periods of data are required for correct identification in this case. This is reasonable, considering the signal is periodic and without noise. We obtain all required information within a few oscillations of the signal.

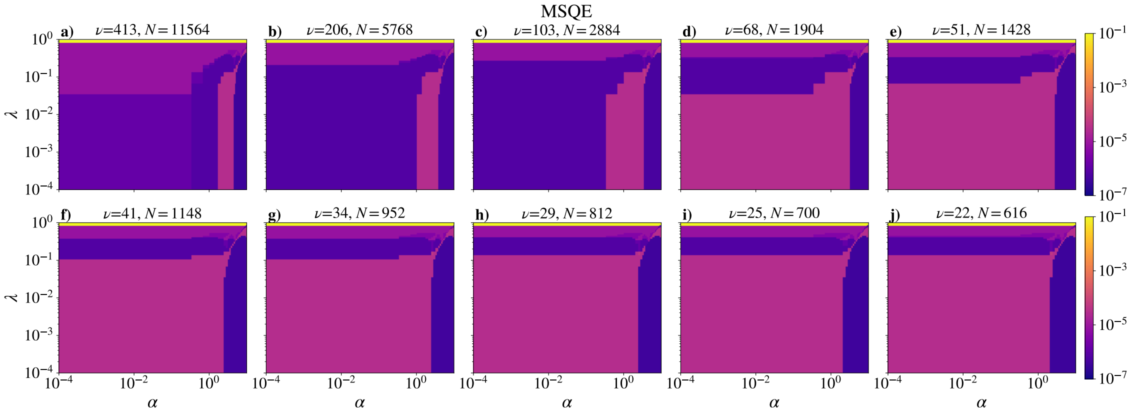


Figure 139: MSQE of the predicted SINDy model on the validation set for varying sampling rates with pre-calculated derivatives. Each subplot is taken for a different sampling rate ν and total number of samples N marked in the subplot title.

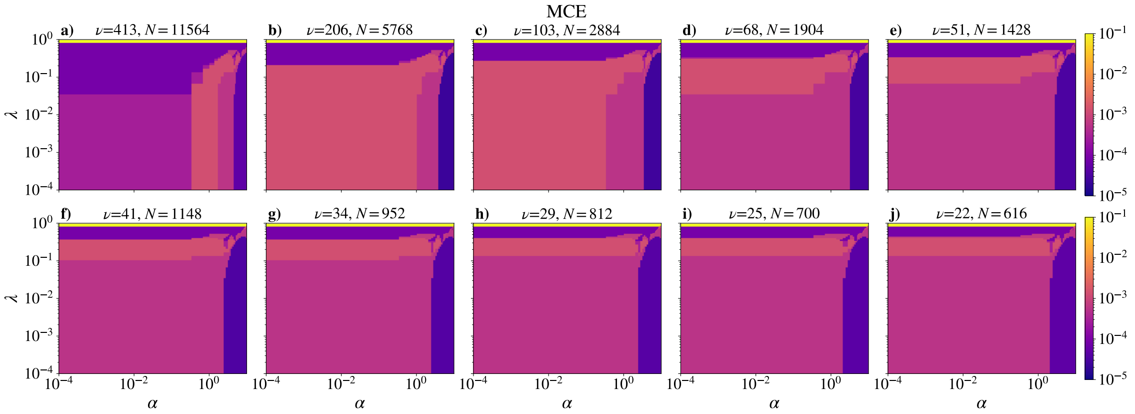


Figure 140: MCE of the predicted SINDy model for varying sampling rates with pre-calculated derivatives. Each subplot is taken for a different sampling rate ν and total number of samples N marked in the subplot title.

8.1.3 MIOSR with convential SINDy

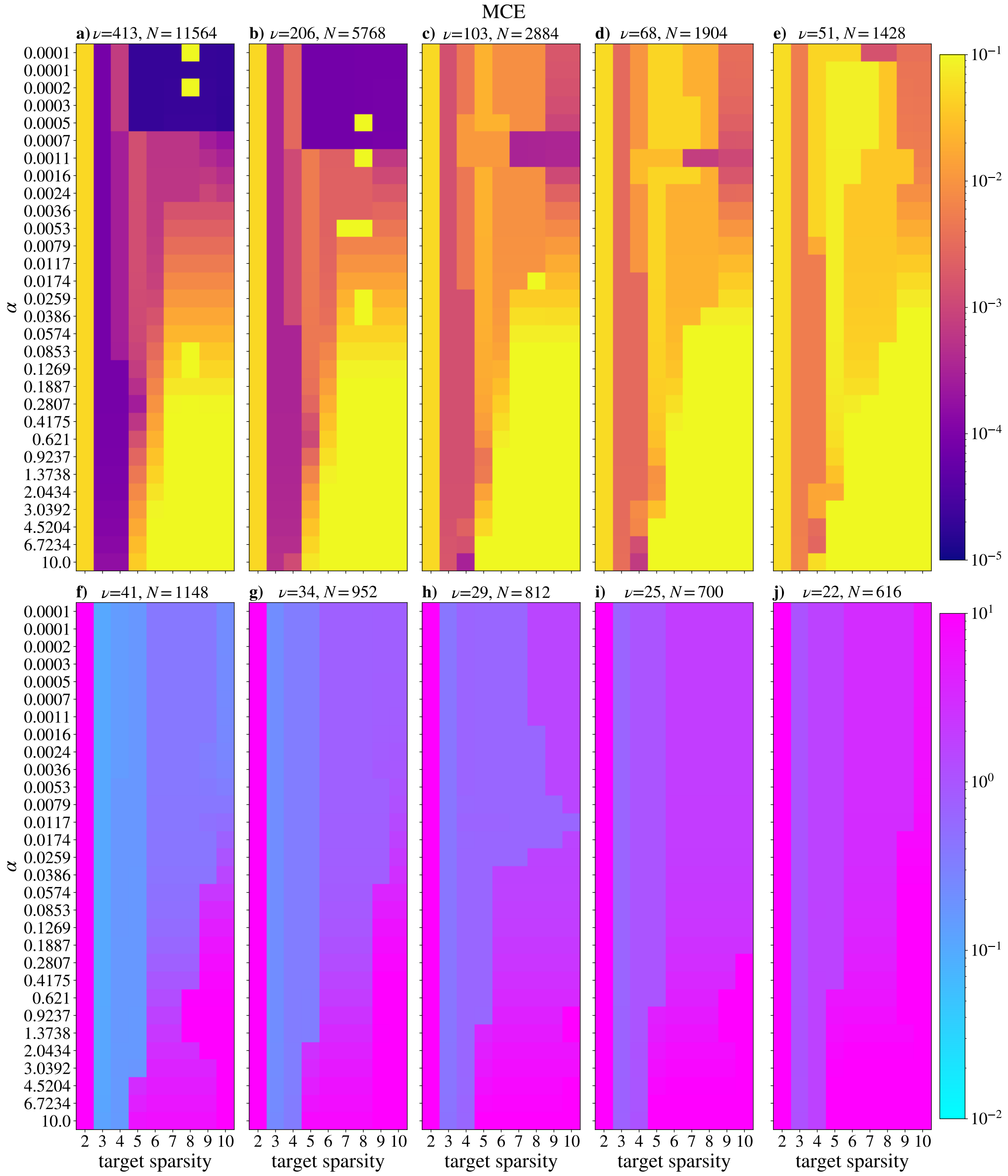


Figure 141: Calculation of the MCE using MIOSR optimiser for varying target sparsity, α and sampling rates. Each box signifies a different sampling rate. The sampling rate ν and total number of samples N are shown in the title of each subplot. The true sparsity of the ANAC model is $k = 3$.

One possible solution for the issues with low-sampling rates that was explored was to employ a different optimiser. A recent benchmarking study by ref. [163] suggests that in noisy studies of a selection of chaotic systems, MIOSR and STLSQ with weak SINDy perform best of the available optimisers. Here we first consider only implementing the MIOSR optimiser. As discussed in previous sections, the main advantage on MIOSR is specifying an allowable sparsity of the final solution (k).

To construct a comparison we implement MIOSR and vary the L2 regularisation and target sparsity for progressively coarser datasets. Here we only show the results for the MCE while using MIOSR in Figure 141. We can see that little improvement in the error is found for differing sampling rates. Regardless, this Figure gives some insight into which parameters provide the best solutions with MIOSR. For high sampling rates, we see that a wide range of target sparsity solutions are possible. While this may be surprising, it is entirely possible for MIOSR to find solutions lower than the specified target sparsity. On the finely sampled data, this is the case with correct model identification still occurring at higher target sparsity. We can also see that for low target sparsities, varying α has little impact on the resulting solution. This is expected, considering for low target sparsities we will generally have problems which are better conditioned and thus benefit less from L2 regularisation.

As the sampling rate decreases, we see that higher target sparsities are no longer favoured and selection of the correct number of terms becomes increasingly important. In some cases, we can obtain reasonable estimates for a target sparsity of 4 by an appropriate selection of α. From these conclusions, we can see that even using a solver with fewer numbers of terms appearing in the solution we still cannot produce accurate results due to the errors from finite differencing.

8.2 Weak SINDy

8.2.1 Weak SINDy as a resolution to sparsely sampled data

The most obvious remedy to poor equation recovery at low sampling rates is to employ a method which avoids taking derivatives entirely. As we aim to identify ODEs which can be written in the weak form, we are able to integrate over all polynomial features in the feature library in time and thus avoid taking derivatives entirely. However, we have found that when using the weak form, proper selection of the size of the integration domain H_{xt} is essential for reliable model discovery. For initial model recovery, we make use of weak SINDy with the STLSQ optimizer.

There have been several remarks in a variety of publications on the selection of the size of the integration domains, the number of integration domains, and their locations in space or time. Discussions on the number of integration domains has been given by [171], noting that increasing the number of integration domains K generally improves the regression ensuring the data is both more diverse and improving robustness due to averaging in the noisy case [124]. On the flip side, when using a high K, domains should be selected in such a way to ensure linear independence of the entries such that integration domains are minimally overlapping. Locations of the sampling domains is of course relevant, in PDE identification sampling of the boundary layer is typically important for the recovery of viscous terms [171]. The size of the integration domain is also important. In the case of periodic signals if the integration domain spans periods of the signal then weak SINDy can act as a low-pass filter [152]. If signals have low-sampling rates (few samples per period) then the integral windows will be unavoidably large, and we contend with a trade-off between noise averaging and high-frequency filtering [170].

Figure 142 shows the calculation of the MSQE over the training data. We see that lower MSQE in general is obtained for lower values of α. This is not unusual as we are calculating an error based on a clean training set, and so these models are either overfitted or benefit from small regularisation due to the clean data. We can see, however, that at the smallest values of α, there are instances where correlated features result in poorer model identification with this becoming more robust at increasing values of α. As α continues to increase beyond $\alpha = 10^{-5}$, the region of acceptable H_{xt} reduces substantially.

To understand why this happens, we also need to consider the MCE in Figure 143. This shows a different result where, even at low values of α, the acceptable region of H_{xt} cannot reasonably exceed one period in the data. As α increases we again achieve more robust model identification with the lowest MCEs centred around $H_{xt} = 0.3\Lambda$ for $\alpha = 1 \times 10^{-4}$. This suggests that at a low α, the problem can be overfitted and still achieve a good MSQE on the training set. Increasing the regularisation then reduces the impact of correlated features showing an acceptable selection for H_{xt}. As α increases to 0.1, the regularisation is too large and the model favours fitting the $L2$ constraint over the data.

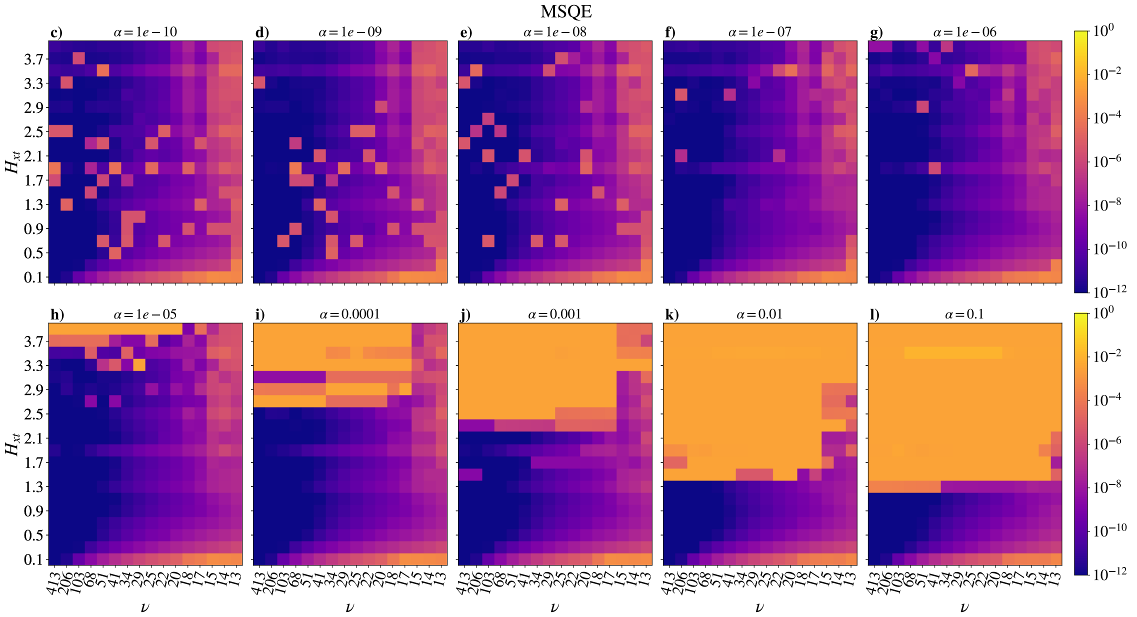


Figure 142: MSQE of the formulated integrated feature library for varying size of the integration domain H_{xt}, the sampling rate ν and the L2 regularisation α with a fixed number of integration domains $K = 10000$. Each box represents a different value of α

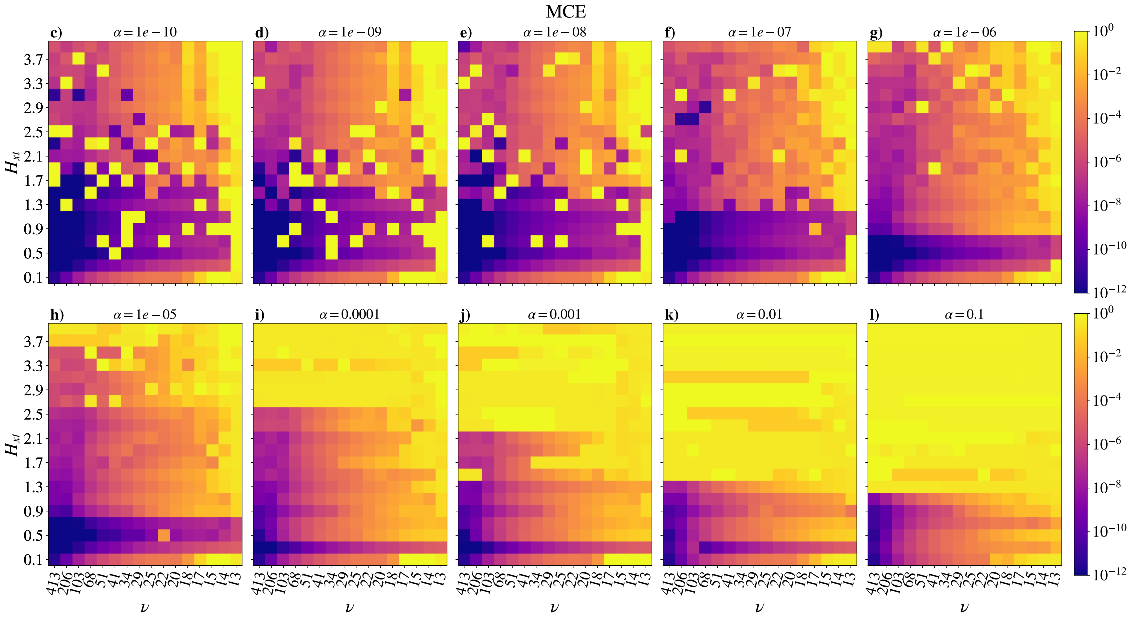


Figure 143: MCE of the formulated integrated feature library for varying size of the integration domain H_{xt}, the sampling rate ν and the L2 regularisation α with a fixed number of integration domains $K = 10000$. Each subplot represents a different value of α.

We also check the sufficiency of the training data length in the noiseless case. Figure 144 shows the MCE while varying the sampling rate and total length of the data. We can see that the main increase in error relates to decreasing the sampling rate in the series, with an upper limit of $\nu \approx 10$ for successful recovery. However, decreasing the sampling rate still causes an increase in coefficient errors possibly related to the loss of high-frequency components in the signal. Little to no benefit is obtained on coefficient estimates by increasing the total training length. This is a result of the data being noiseless and oscillatory. Adding additional data does not explore more of the phase space due to the nature of the system and does not average over any noise in the system.

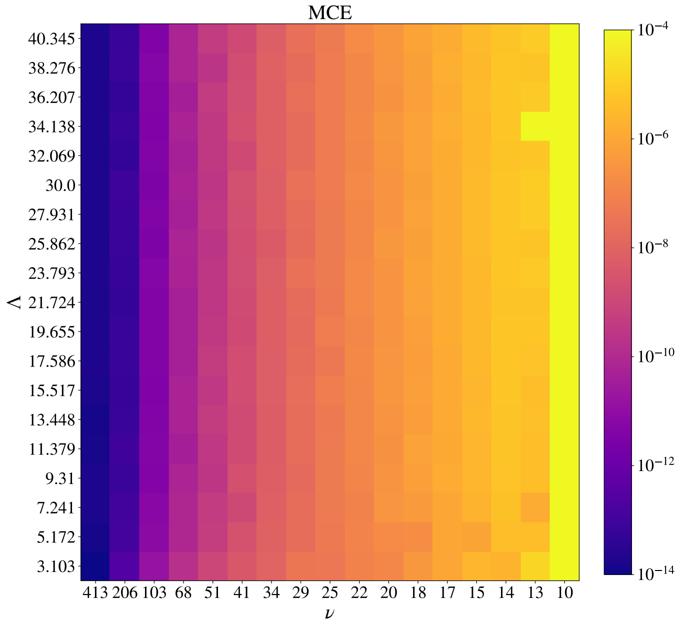


Figure 144: Comparison of the MCE while varying the total length of training data and sampling rate. Parameters fixed at: $K = 10000$, $H_{xt} = 0.3\Lambda$, $\alpha = 1 \times 10^{-4}$ and $\lambda = 0.1$.

8.2.2 Weak SINDy with Noise

As weak SINDY with STLSQ is capable of recovering accurate estimates of the underlying governing equations for the desired sampling rates, we now look at recovery of the equations in the noisy case. Unfortunately with noise added and weak STLSQ SINDy, many of the above results deteriorate (not shown). This is simply due to correlated input features with periodic signals, as STLSQ has no way of limiting the resulting sparsity of the final solution. As a consequence, many non-zero terms can be included which effectively cancel out.

In an attempt to reduce the correlation issue, we make use of the MIOSR optimiser with weak SINDy, again limiting the total sparsity to 3 (3 non-zero coefficients). The results for the MCE are shown in Figure 145 for varying α. By restricting the resulting model to be sparse, we vastly improve the identification of the underlying model in higher levels of noise. For additive noise of a degree $\delta_{\text{obs}} \leq 0.1$, the coefficients are accurately determined for all sampling rates. However, for higher degrees of noise at $\delta_{\text{obs}} = 0.5$ and the sampling rates less than $\nu \approx 50$ we see that incorrect models have been found.

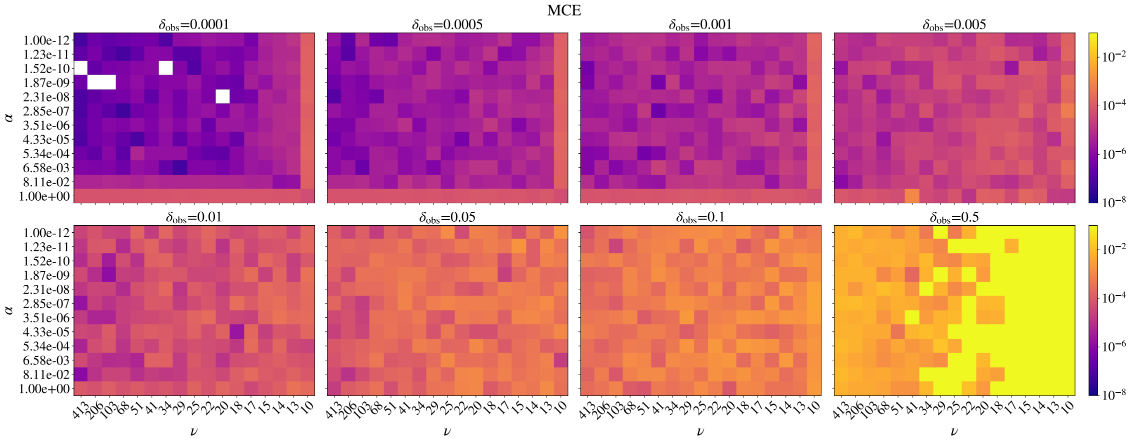


Figure 145: MCE when applying weak SINDy with MIOSR optimiser and an ideal fixed target sparsity of 3. In each subplot the L2 regularisation α and sampling rate ν are varied for varying degrees of noise, with standard deviations listed in the subplot titles.

8.2.3 Constraints for the ANAC model

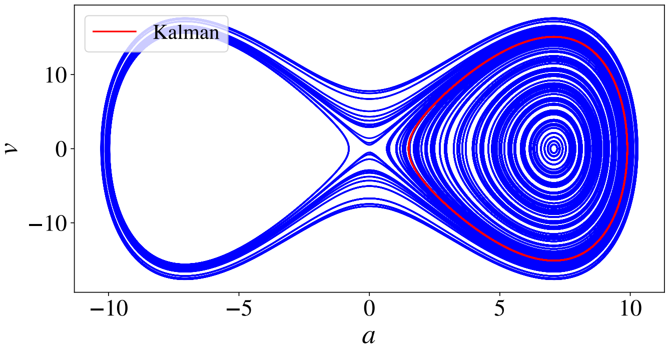


Figure 146: The phase-space of the ANAC4 model for several initial conditions shown in blue with the training data shown in red. The red data is the equivalent to the data considered by ref. [115].

So far we have only discussed cases where feature selection is based purely on time series data. An advantage can be gained from considering constrained models which obey some physical principle, thus reducing the allowable search space. The first case we consider are Hamiltonian constraints for when the system is energy-preserving. This is relevant in, for example, ideal MHD, and for the ANAC model which is energy-preserving. For a 2D dynamical system

$$\dot{a} = f(a, v), \tag{8.4}$$

$$\dot{v} = g(a, v) \tag{8.5}$$

to be Hamiltonian, we only require that

$$\partial_a f(a, v) + \partial_v g(a, v) = 0. \tag{8.6}$$

For a third-order polynomial SINDy library of the form

$$\dot{a} = c_0 + c_1 a + c_2 v + c_3 a^2 + c_4 v^2 + c_5 av + c_6 a^3 + c_7 v^3 + c_8 av^2 + c_9 a^2 v,$$
$$\dot{v} = d_0 + d_1 a + d_2 v + d_3 a^2 + d_4 v^2 + d_5 av + d_6 a^3 + d_7 v^3 + d_8 av^2 + d_9 a^2 v,$$

we require the following constraints

$$
\begin{aligned}
c_1 = -d_2, \qquad 2c_3 = -d_5, \qquad c_5 = -2d_4, \\
3c_6 = -d_9 \qquad c_8 = -3d_7, \qquad c_9 = -d_8.
\end{aligned}
$$

If these constraints are directly implemented, then the resulting models will be energy preserving. In ref. [72] the model is also derived under the assumption of the symmetry $a \to -a$ and $v \to -v$. For the third-order polynomial library, this gives many different additional constraints. For $\dot{a}$ we have

$$c_0 = 0, \quad c_3 = 0, \quad c_4 = 0, \quad c_5 = 0,$$

and similarly for $\dot{v}$

$$d_0 = 0, \quad d_3 = 0, \quad d_4 = 0, \quad d_5 = 0.$$

The symmetry constraint provides new information if we only have the training data shown in Figure 146. The training data we have used until now only shows oscillations in one part of the phase space (one side of the potential well) and so it is possible to fit oscillating models which do not generalise to the entire phase-space. If all the constraints are applied, the resulting model will be of the form

$$
\begin{aligned}
\dot{a} &= c_1 a + c_2 v + c_6 a^3 + c_7 v^3 + c_8 a v^2 + c_9 a^2 v, \\
\dot{v} &= d_1 a - c_1 v + d_6 a^3 - \tfrac{1}{3} c_8 v^3 - c_9 a v^2 - 3 c_6 a^2 v.
\end{aligned}
$$

This reduces the number of unknowns from 20, to 8.

On occasion, SINDy models have been derived by considering apparent symmetries in the phase space by looking at plots of the trajectories [148] like those shown in Figure 146. If we followed such a process here, we would end up considering the symmetry $a \to -a$ and $v \to v$ (shown in red on the training data) which is not the correct symmetry of the entire system.

8.2.4 Applying constraints to the weak form

Given the desired properties of our system, we can reapply the assessment with added noise when either Hamiltonian or symmetry constraints are applied. Inclusion of the Hamiltonian constraints in Figure 147 does not appear to lead to any substantial improvement in the resulting model fits. This is simply because the constraints supply no information about the terms we would actually hope to retain in our resulting model. That is, when the correct terms are identified there, any additional terms are set to zero anyway and the Hamiltonian constraints provide no relation between the resulting non-zero coefficients c_2, d_1 and c_6. As a result of the constraint on the overall sparsity, the only terms which are actively fitted are terms not related to these particular set of constraints.

For the symmetry constraints in Figure 148 we instead see a good improvement as these provide new information to the SINDy model. As the training trajectory is bounded within one side of the potential well only, there is no information on the full phase space and therefore the full symmetry of the potential well. For δ_{obs} we are able to still identify models with the correct coefficients active at sampling rates of around $\nu = 12$.

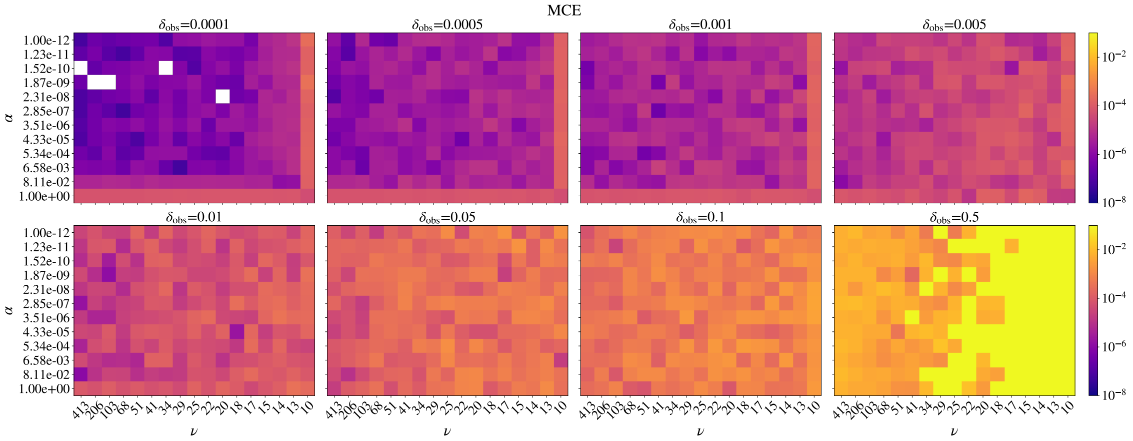


Figure 147: Applying weak SINDy with MIOSR optimiser and an ideal fixed target sparsity of 3 with Hamiltonian constraints. In each subplot the L2 regularisation α and sampling rate ν are varied for varying degrees of noise, with standard deviations listed in the subplot titles.

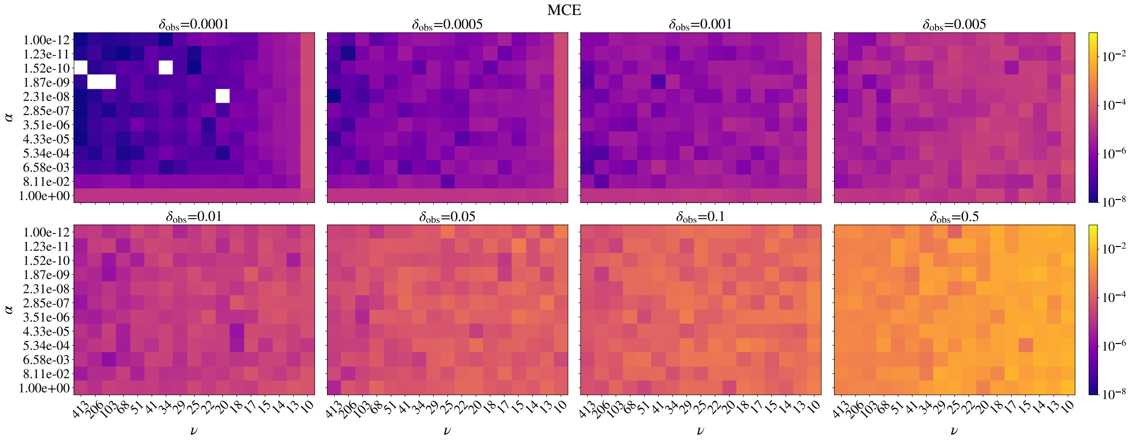


Figure 148: Applying weak SINDy with MIOSR optimiser and an ideal fixed target sparsity of 3 with symmetry constraints. In each subplot the L2 regularisation α and sampling rate ν are varied for varying degrees of noise, with standard deviations listed in the subplot titles.

8.2.5 Redundant training trajectories

The trajectories of the ANAC model can be related to the motion of a particle confined to a potential well, the potential well being plotted in Figure 18. For small enough amplitudes of a, the motion of the particle is confined in either the left or right potential wells. If the training data is only taken on one side of these potential wells, we will often identify incorrect models irrespective of the sampling rates and degrees of noise added. If we train from a trajectory close to the fixed point at the right potential well with initial conditions $(a_0, v_0) = (7.03, 0.01)$ with no noise and only Hamiltonian constraints, the resulting model identified has the form

$$\dot{a} = 1.000v, \tag{8.7}$$

$$\dot{v} = 2.828aa - 0.400aaa \tag{8.8}$$

showing incorrect terms being fitted (no linear a term). We emphasise that symmetry constraints have not been applied here. The above model estimates the right fixed point of the phase space as $a^* \approx 7.07$ and the fixed point at $a = 0$ but not the left fixed point. While this behaviour is exacerbated for trajectories closer to the fixed point, downsampling and noise can also cause loss of information in the phase space and result in solutions like this if the trajectories are trapped on one side of the potential well. This explicitly shows why the symmetry constraints are beneficial. Application of the symmetry constraints eliminates this model as a viable candidate.

8.3 Comparison between EnKF and SINDy for parameter inversion

Method	$\bar{\mu}$	$\bar{\sigma}$	$\delta\mu$	$\delta\sigma$	Failed Runs
EnKF	10.05735	-0.09985	0.35824	0.01871	2
Constrained EnKF	10.15952	-0.09775	0.02299	0.00021	0
SINDy	9.942428	-0.098064	0.07866	0.00162	N/A

Table 14: Comparison of identified parameters between EnKF and SINDy in the two-state measurement case.

Comparing both SINDy and EnKF on an equal footing is challenging because many different factors can impact the performance of each method. For EnKF, correct selection of the measurement noise, suitable initial guesses and appropriate choice of priors will impact the quality of the resulting fit. For SINDy the allowed sparsity, numbers of features included in the library, and different learning parameters can also produce differing answers. We attempt to select values for each of these methods which represent best case scenarios for both methods.

For the data, we adopt a sampling rate of $\nu \approx 12$ with additive noise with standard deviation $\delta_{\text{obs}} = 0.5$. The data is simulated for 35 total periods with both a and $\dot{a}$ being treated as observable. For SINDy, we make use of both Hamiltonian and symmetry constraints with the MIOSR optimiser and a cubic polynomial library. The optimiser is set to a fixed target sparsity of 3 with a low regularisation of $\alpha = 1 \times 10^{-12}$, $K = 10000$ and $H_{xt} = 0.3\Lambda$. In EnKF, we set the perturbations to the initial true initial conditions to be drawn from uniform distributions with $\tilde{a}_0 \in [-4, 4]$, $\tilde{v}_0 \in [-4, 4]$, $\tilde{\mu}_0 \in [0, 20]$ and $\tilde{\sigma}_0 \in [-2, 0]$. The initial covariance matrix is set to roughly reflect the uncertainty in the initial condition with $\boldsymbol{P} = \text{diag}[2, 2, 3, 1]$ and the measurement uncertainty matrix is set to $\boldsymbol{R} = \text{diag}[0.5^2, 0.5^2]$. We set the number of ensemble members to be fixed at $N = 30$ throughout.

To compare the results, we compute the means and standard deviations of the estimates of μ and σ for 100 random initial conditions in the EnKF method. For SINDy, we compute the means and standard deviations from 100 models which are fit with a different seed generating the noise distribution each time. We also compute two cases with EnKF, unconstrained and constraining $\sigma < -1 \times 10^{-16}$. When using constrained EnKF, we set $\boldsymbol{Q} = \text{diag}[0, 0, 1 \times 10^{-3}, 1 \times 10^{-4}]$ and $\boldsymbol{Q} = \boldsymbol{0}$ otherwise. The choice of $\boldsymbol{Q}$ in the constrained case is based on the previous observation that constrained σ artificially reduces uncertainty and results in poor parameter convergence. We therefore generate an uncertainty in σ which prevents collapse of the ensemble members at the constraint bound.

Table 14 shows the comparison between SINDy and EnKF. For conventional EnKF, we typically achieve good convergence albeit some runs diverge because $\sigma > 0$. The poor convergence of these runs causes a slightly larger standard deviation in the estimate of the parameter values. For constrained EnKF, there are no diverging runs, and all runs converge well close to the true parameter values. There is a markedly lower deviation in σ for this case due to the constraints reducing the uncertainty. SINDy also produces accurate estimates of the true model coefficients, identifying the correct active terms in all cases.

All methods compare well with one another in the case where full-state measurements are available.

Compared to the unconstrained case, SINDy is less likely to converge to incorrect parameter estimates whereas EnKF on occasion does. However, both methods solve different optimisation problems. For EnKF the state must also be estimated as well as the parameters and with SINDy, there are a total of 8 possible unknown coefficients. Even though SINDy has 8 possible coefficients, the restriction of the total sparsity to be three makes this fitting procedure feasible.

The advantage of SINDy lies in the fact that no knowledge of the parameters is needed beforehand in the form of an initial guess. For library terms, while some assumption must be made on the feature library, more general feature libraries can be given as SINDy does not integrate the resulting models. Further, no estimate of the degree of noise needs to be given to make estimates of the parameters. However, in reality full-state measurement cases are unlikely or at least not guaranteed when using tokamak data. With the ANAC model, it is proposed that the mode a can be related to the magnetic field and therefore fits would be performed to Mirnov signals (see refs. [84, 115]). Mirnov signals measure the rate of change in the magnetic field and therefore the only observable variable is $\dot{a}$. In this sense, EnKF has a natural extension for single measurement cases which is not the case for SINDy. Further, to get comparable performance SINDy must be used with quite a restrictive library with relatively few unknowns. Without this, identification almost totally fails in cases where EnKF does not struggle at all. In this sense SINDy cannot be used with an arbitrarily constructed library, some thought must be given to the physical constraints the library must obey.

8.4 Concluding remarks

In this chapter we have outlined an approach for SINDy when applied to the ANAC model. Similarly to the conclusions drawn with the Knobloch system of equations in §4, we see that to apply SINDy to noisy, sparsely sampled signals the weak form is essential. While filtering approaches can be used to reduce noise, avoiding the need to find derivatives of noisy data is the simpler approach. We also discussed implementation of constraints which have been developed for the ANAC model by ref. [72] and how they applied to SINDy models. For the applied symmetry constraints, these offer the most improvement when knowledge of the complete symmetry is not shown in the training data. Implementations of Hamiltonian constraints do not offer much improvement but at least offer a guarantee of models obeying physical principles. Hamiltonian constraints would be relevant where ideal MHD is relevant, as is often the case in understanding instabilities in tokamaks [34].

Finally, we compared the performance of EnKF to SINDy and attempted to use both methods in an optimal setting. For full-state measurements, both approaches perform very well, though creating an equal comparison is challenging as initial estimates of the parameters and priors must be given to EnKF which is not the case for SINDy. In this sense, SINDy offers an advantage as no prior knowledge of the parameters is required. It also considers a more general fitting process where any model that obeys Hamiltonian constraints and the outlined symmetries can be selected.

The main difficulty in using SINDy comes from the lack of full-state measurements for which EnKF is capable of addressing. In reality, Mirnov coils only measure the change in magnetic field, which has been likened to the evolution of $\dot{a}$ in the ANAC and ANAET models. For SINDy, this can be beneficial as we now only must estimate the derivatives for $\dot{v} = \ddot{a}$. On the other hand, we must integrate $\dot{a}$ to obtain a, implying that we will only know a up to a constant. We can regard this as the change of variables $a \to \bar{a} + a_0$ where a_0 is a constant of integration. Using this change of variables in the ANAC model gives the result

$$\dot{\bar{a}} = v, \tag{8.9}$$

$$\dot{v} = (\mu a_0 + 6\sigma a_0^3) + (\mu + 6\sigma a_0^2)\bar{a} + 6\sigma a_0 \bar{a}^2 + 2\sigma \bar{a}^3. \tag{8.10}$$

This is now the system SINDy will attempt to fit and depends on the initial condition a_0. We can further see that the resulting system obeys the symmetry constraint $a \to -a$ and $v \to -v$, but not the symmetry $\bar{a} \to -\bar{a}$ and $v \to -v$. While we describe this for a constant of integration, a similar result will occur for any non-zero constant noise. If constraints are to be applied with SINDy in tokamak data, a process would have to be developed which could address this.

We also must pay attention to the quality of the reconstruction of a from integration. Figure 149 shows a comparison of the results of using cumulative trapezoidal integration on 3 different sampling rates with comparisons of the results in the phase space. For the sampling rate of $\nu = 413$, integration of the trajectory produces the correct orientation of the result in the phase space and represents the training data used to identify the SINDy model in the previous section. The only difference here corresponds to a translation from the unknown constant of integration. For higher sampling rates, we can see that poorer sampling of the right-hand side of the phase space results in an artificial skew introduced in the phase space. We see that there is a much larger change in amplitude in this section of the phase space, and thus a poorer approximation of the integral. We can also note that this skewing of the results has a significant impact on the results at sampling rates much higher than we aim to reach. The skewing will result of the trajectories in the phase space then produces a trajectory which no longer exhibits the desired symmetry in the phase space.

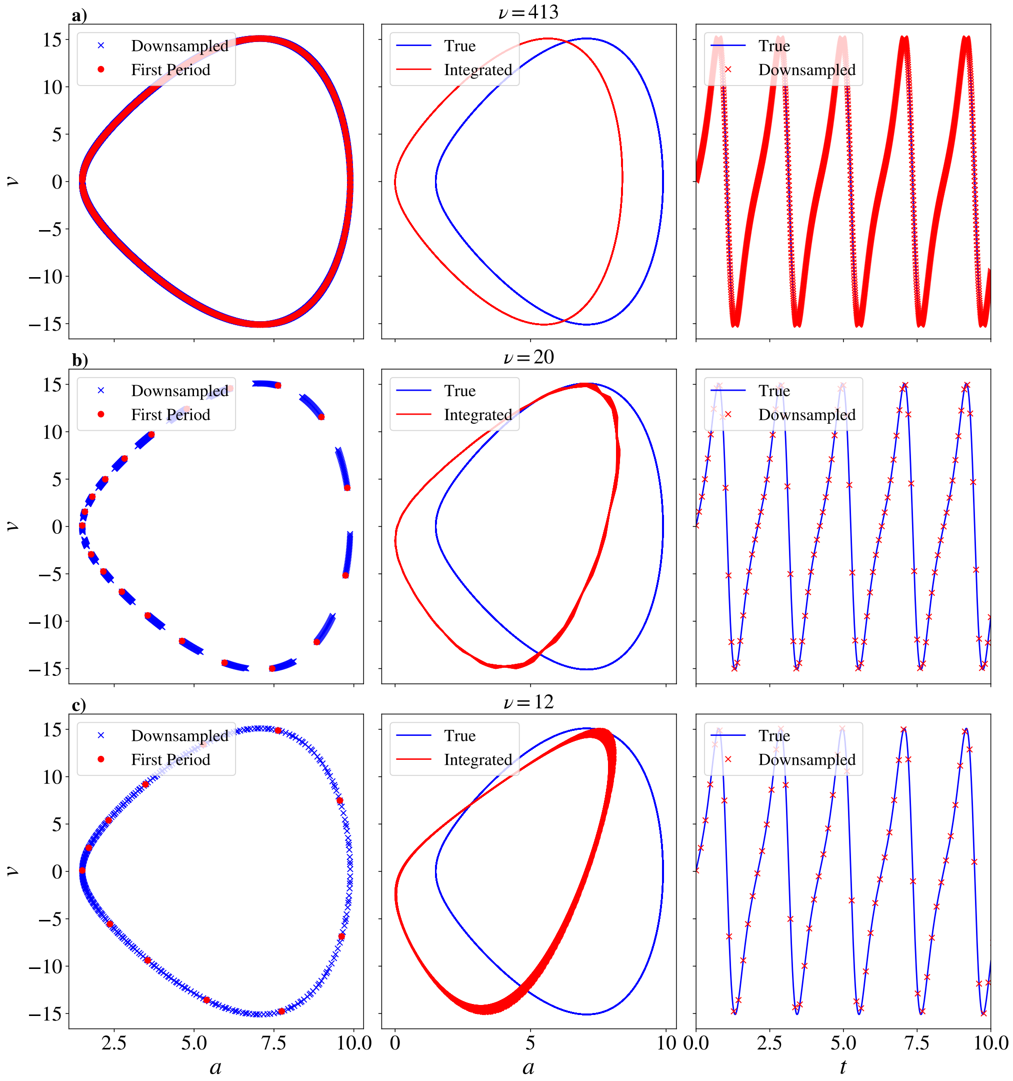


Figure 149: Comparison of the results of cumulative trapezoidal integration for noiseless downsampled trajectories. Each row corresponds to the specified sampling rate, showing the phase space of the downsampled trajectory in the left column, the comparison of the true and integrated solutions in the middle column and an example of the downsampled trajectories in the right-hand column. Each row is titled with the sampling rate ν and we emphasise that this is different from $\dot{a} = v$ shown in the $y-$label.

9 Conclusions

In this chapter we summarise the work presented in this dissertation, namely the results given in §4, §5, §6, §7 and §8. We will conclude with a discussion on avenues of future work.

9.1 Summary of results and conclusions

SINDy with magnetoconvection

We first began by benchmarking SINDy to a weakly nonlinear model representative of the magnetoconvection PDE given by equations (4.45)-(4.46). The weakly nonlinear model provides a suitable testing scenario for SINDy because it exhibits a wide variety of behaviours and comes from resistive MHD, physics commonly incorporated in tokamak models (e.g, ref. [131]). By validating SINDy's performance on the weakly nonlinear model, we could assess the performance in a range of cases such as noise robustness, impacts of nonlinearity, choice of optimiser, and sensitivities to data length and sampling rates. It was found that a conventional implementation of SINDy following the original method outlined in ref. [97] showed several instances of poor noise robustness. This manifested in multiple scenarios, such as cases with correlated input time series and poorer sampling rates relative to the fastest scales in the system. Even in cases where the feature library consists of only second-order polynomial terms, if high quality derivative estimates are not provided, equation recovery is poor. In this sense, SINDy does not function as a method capable of equation recovery with exhaustive libraries in highly generalised cases. Instead, much more care must be taken when constructing the library and it is not an out of the box approach.

Alleviation of issues introduced by both noise and sampling rates is achieved by using the weak form of SINDy [142]. It was shown that the performance of the weak form far exceeds that of the conventional approach, particularly with low sampling rates and noise. The major caveat to this comes from selection of integration domain windows. We find that the window size needs to limited to less than approximately one characteristic period of the system. Such a timescale could easily be determined by, for example, a fast-Fourier transform. The weak form would present issues in multiscale systems where the window size is limited by the fastest scales in the system. Multiscale systems that are forced to have smaller integration windows will then provide a lower degree of noise averaging.

We finally considered the application of the weak form of SINDy with constraints derived from conditions satisfied by the weakly nonlinear model. These constraints were implemented with the MIOSR optimiser which satisfies specified constraints exactly [160]. While this approach introduces normalisation issues (which can be easily treated), it ultimately results in models which are guaranteed to obey the outlined constraints. An example considered with noisy, sparsely sampled training data showed that the constrained model produced a better representation of the dynamics than an unconstrained one, despite it not recovering the true underlying system.

SINDy with magnetconvection PDE data

In this section, we considered the application of SINDy with POD modes derived from magnetoconvection simulations of the full 2D PDE system. The application with POD modes was considered because close to bifurcation we expected the derived weakly nonlinear system to be valid and the POD modes to represent Fourier modes i.e. the basis decomposition used in deriving the weakly nonlinear model.

We first considered the application of SINDy to POD datasets at different Rayleigh numbers exhibiting overstable oscillations. We noted that errors introduced by the POD both prevented recovery of the weakly nonlinear model close to the onset of overstable oscillations and impeded selection of sparse models using STLSQ. For this reason, we opted to use MIOSR optimiser as it restricts the number of non-zero coefficients allowed in the resulting models. By restricting k we can generate a curve of models for varying numbers of non-zero coefficients and observe the Pareto front which is not always possible when using STLSQ.

Results were then presented for each Rayleigh number showing the MSQE errors for models which implemented no constraints, symmetry constraints, diffusive constraints or symmetry and diffusive constraints for both second and third-order polynomial libraries. For the second-order polynomial library, these results showed sparsities of $k = 10 - 14$ were sufficient to obtain low MSQEs for $R \leq 3000$, but at higher Rayleigh numbers, denser models were required to achieve low MSQE. The same trend was seen with the third-order polynomial library, though for Rayleigh numbers $R > 3000$ the third-order polynomial library required fewer terms than the second-order polynomial library. In either case, close to the onset of overstable oscillations, model sparsities lower than $k = 14$ corresponding to the number of non-zero coefficients in the weakly nonlinear model are favoured. Again, for the second and third-order polynomial library, MSQEs were not appreciably lower as a result of including constraints. Despite this, the resulting models are guaranteed to obey the constraints which is not always true in the unconstrained case.

We then considered how models from different Rayleigh numbers could be parameterised. Fitting a single model over POD time series taken from different Rayleigh numbers successfully parameterised trajectories within the training range and predicted a transition to no convection as the Rayleigh number decreased. However, as R was increased the model failed to predict a bifurcation to steady convection. The issues arose from variation of the POD basis at different Rayleigh numbers. Study of a fully constrained model derived at fixed Rayleigh number $R = 6000$ with $k = 17$ and a third-order polynomial library showed both a bifurcation to no convection as R was decreased and bifurcation to steady convection as R increased.

Finally, we considered deriving simplified models from POD time series taken at parameter values exhibiting chaotic convection. Three different scoring methods were considered: the AIC_c score, the KL divergence, and the maximum number of predicted Lyapunov times n_λ. By making use of TPEs, we performed model selection using these metrics while searching over different allowable sparsities for each equation. TPEs offer an approach for performing a more robust selection of hyperparameters than standard grid-search methods. We found that model selection metrics which required integration (the KL and n_λ scores) favoured sparser models compared to the AIC_c score. Both the KL and n_λ scoring methods produced models capable of reproducing both short and long term dynamics of the system. We then concluded by presenting a model of 15 non-zero terms which performs as well as the weakly nonlinear model but is constructed only from four POD modes.

A data assimilation approach with the ensemble Kalman filter

We began by introducing a data assimilation approach called the ensemble Kalman filter. Several extensions of this method were discussed and their relevance to the ultimate end goal of applying this method to experimental tokamak time series. We first discussed implementation of a new code versus an older pre-existing code which was implemented by [115]. We then applied EnKF to the ANAC model to assess the performance of EnKF with parameter estimation and poor initial guesses. We addressed concerns raised by ref. [106] which essentially boil down to the performance of EnKF being sensitive to initial conditions with the ANAC model. We identified two reasons for this, neither of which appear to be discussed by ref. [106]. Both reasons connect to an appropriate choice of the ensemble spread or initial ensemble uncertainty denoted $\boldsymbol{P}_0$. The initial ensemble uncertainty must be: 1. reflective of the uncertainty in the initial guesses and 2. contain physically plausible realisations of ensemble members. While the work of [106] introduces an interesting optimisation approach for improving initial guesses, we find robust performance of EnKF in the same range of initial guesses without the need for this optimisation. By adding a simple constraint method on ensemble members, we also introduced an approach which allows ensemble members to only be drawn from physically realistic parameter regions. This can, however, cause an artificial collapse of the ensemble at constraint bounds.

We also extended EnKF to single measurement cases of the ANAC model where only measurements of $\dot{a}$ are available. This is more relevant to experiments, where only a single Mirnov measurement will be considered. This means that EnKF must infer the unobserved variable a, as well as the unknown parameters of the ANAC model. We showed that EnKF still coped with this case well and was robust to errors in the estimate of the observational error matrix $\boldsymbol{R}$. For the ANAC model, this represents a complete validation that can realistically be performed without the application to experimental data.

EnKF and that ANAET model - progressing to experimental conditions

Following validation of EnKF against pre-existing work with the ANAC model, we then extended the application to the ANAET model, which presents a realistic comparison to experiment. The ANAET model has additional challenges when implemented with EnKF namely: it is a multiscale system so low sampling rates can produce poor sampling of the fastest timescales, there are now up to 8 unknown coefficients to be determined, and single-measurement cases would have 2 hidden states along with unknown parameter. We validated the performance of EnKF in both full-state measurement cases, double-state measurement cases (with $\dot{a}$ and b) and single-state measurement cases (a only). In all cases, EnKF is remarkably robust when initial conditions are reasonably close to the true values, but we note that arbitrary initial ensembles will often result in divergence of the filter. One reason for this is that the high-order diffusive term of the ANAET model has the form $\mu_6 a^6 \dot{a}$ and if $\mu_6 > 0$ this becomes destabilising. There are, of course, other ways in which ensemble members can become unstable.

We finally considered extensions of EnKF with the ANAET model when a stochastic integration scheme was used to generate the observations. As discussed in §2.4.3, stochastic integrations of the ANAET model exhibit several features from experiment such as aperiodic sawtooth crashing. We first showed that if observations are generated from stochastic integrations of the ANAET model and exhibit aperiodic sawtoothing, EnKF does not successfully converge. This is because that parameters are modelled as constants with no error, so it is not possible for EnKF to predict quasiperiodic sawtoothing without varying the parameters. We do suggest that in reality, parameters should be allowed to vary but this introduces many further difficulties as we do not know beforehand what the parameter values should be. By using stochastic observations, we highlight a case where we restrict the sampling rates to represent an experimental measurement from MAST-U with comparable degrees of noise and show that despite poor initial estimates of parameters we still have good convergence. This is achieved by placing constraints on the diffusive term so that it remains diffusive throughout assimilation. We also introduced a method to estimate the observational noise by taking the FFT and removing high frequency components. This approach provided good estimates for the noise variance added to the observational dataset and can be used to estimate $\boldsymbol{R}$.

SINDy and the ANAC model

Here we presented a brief outline of constraints and extensions required for SINDy to be used with the ANAC model and under similar conditions to those presented with EnKF in §6. In cases where low sampling rates and noise are present, the weak form is essential as expected from §4. By including symmetry and Hamiltonian constraints, we improved the robustness of SINDy to both noise and low sampling rates, noting that this improvement primarily results from the inclusion of symmetry constraints. We then compared SINDy and EnKF for estimating parameter values in the ANAC model when the training data was both noisy and poorly sampled. Both SINDy and EnKF performed well when estimating the parameters despite the high degrees of noise. In all cases the mean estimate of the parameter is within 1% of the true parameter value.

While the performance of SINDy compares favourably in the full state measurement case, there are some considerations. First we are only considering two input variables a and $\dot{a}$ which substantially reduces library correlations. Second, following constraints the library of terms only consists of eight unknowns and is a far cry from the generalised library search of SINDy that was presented originally in [97]. Finally, MIOSR is restricted to only 3 library terms meaning that the final regression is only performed on 3 library terms, not the full library of features. We also must consider the difficulties of applying SINDy with partial measurements. Work presented in this section showed that downsampling of the input time series resulted in poor estimates of the integrals.

9.2 Future work

Applications to tokamak diagnostics

Regarding EnKF, the next step is the application to experimental time series using the diagnostics listed in §2.3. Preliminary matching of EnKF to experimental measurements taken on the MAST-U tokamak was performed using the ANAET model as a space-state model. While this work was not completed due to time constraints, we briefly discuss some results and their consequences for future work. Figure 150 presents an overview of the results of using a soft X-ray measurements as observations for b and Mirnov measurement for observations of $\dot{a}$. The Figure shows the entire range of data trained on, which contains 2 sawtooth crashes located at $t_1 \approx 650$ and $t_2 \approx 1600$.

From Figure 150, several issues are immediately apparent. The data used to train here is only a section of of the entire shot (MAST-U shot 29880), but the chosen Mirnov signal is very noisy outside of this range and often completely unresolved, an issue that was displayed previously in Figure 135. Other scaling issues present themselves. During each ramp phase of the sawtooth, the oscillations often vary substantially in amplitude. The envelope of amplitude of these oscillations also tends to change after a sawtooth crash. This is not a feature that is currently represented when using constant parameters in the ANAET model with EnKF. For the results presented here, this is absorbed by using model error in the observation variables. If the error is allowed to be sufficiently large, it is possible for EnKF to cope but it is not clear how large the model error is allowed to be. Other scaling issues manifest with the spiking behaviour in the Mirnov signals themselves. As was noted previously in Figure 135, the amplitude of the Mirnov signals at many sawtooth crashes can vary substantially. At the second sawtooth crash in this shot, there is a gradual increase in the amplitude of the Mirnov signal and correspondingly the temperature has grown larger than at the previous crash. However, the term in the ANAET model which drives spiking behaviour is of the form ba^3 which grows larger when the amplitude of Mirnov signal grows large, causing the divergence of EnKF shown in bottom of Figure 150. Finally, as discussed in Figure 135 many of the observed spikes in Mirnov signals at sawtooth crashes are neither consistent in behaviour nor occur at the same amplitude. The difference in spiking presents further challenges for a constant parameter model, as the amplitude of spiking tends to be fixed within the first sawtooth crash. Other challenges remain when selecting appropriate measurements as there is a wide selection of shots available within MAST-U each containing many magnetic field measurements. Choosing appropriate data is a difficult task which likely requires someone familiar with shot logs.

On a more positive note, the observations in the selected diagnostics are both concurrent and well sampled, containing around $\nu \approx 60$ samples per fast oscillation. With a sufficiently large model error, EnKF is capable of at least tracking the observations shown in Figure 150. A closer view during a ramp phase in Figure 151 shows that EnKF is performing very well during oscillatory sections of the data. Further, some hope is given in that the selection of many parameters for the ANAET model is not arbitrary. For example, the coefficient γ_1 corresponding to the frequency and amplitude of oscillation of the signal during the ramp phase can be chosen so that $\gamma_1 \approx 1$ with an appropriate rescaling of the data. We can also remark that EnKF is successfully tracking observations despite the outlined challenges in the previous paragraph.

Further applications of data assimilation approaches could benefit by instead using a particle filter approach as has been explored by ref. [162]. In EnKF, each ensemble member or particle is equally weighted and this causes issues when some ensemble members are either non-physical or become unstable. For particle filters, weights are assigned to particles depending on their likelihood of representing the data [147]. For this reason, non-physical members can be assigned zero weights and may allow particle filters to be used with larger prior uncertainties. For the ANAET model, we have little knowledge beforehand of what many of the parameters should be and using larger priors could help combat this.

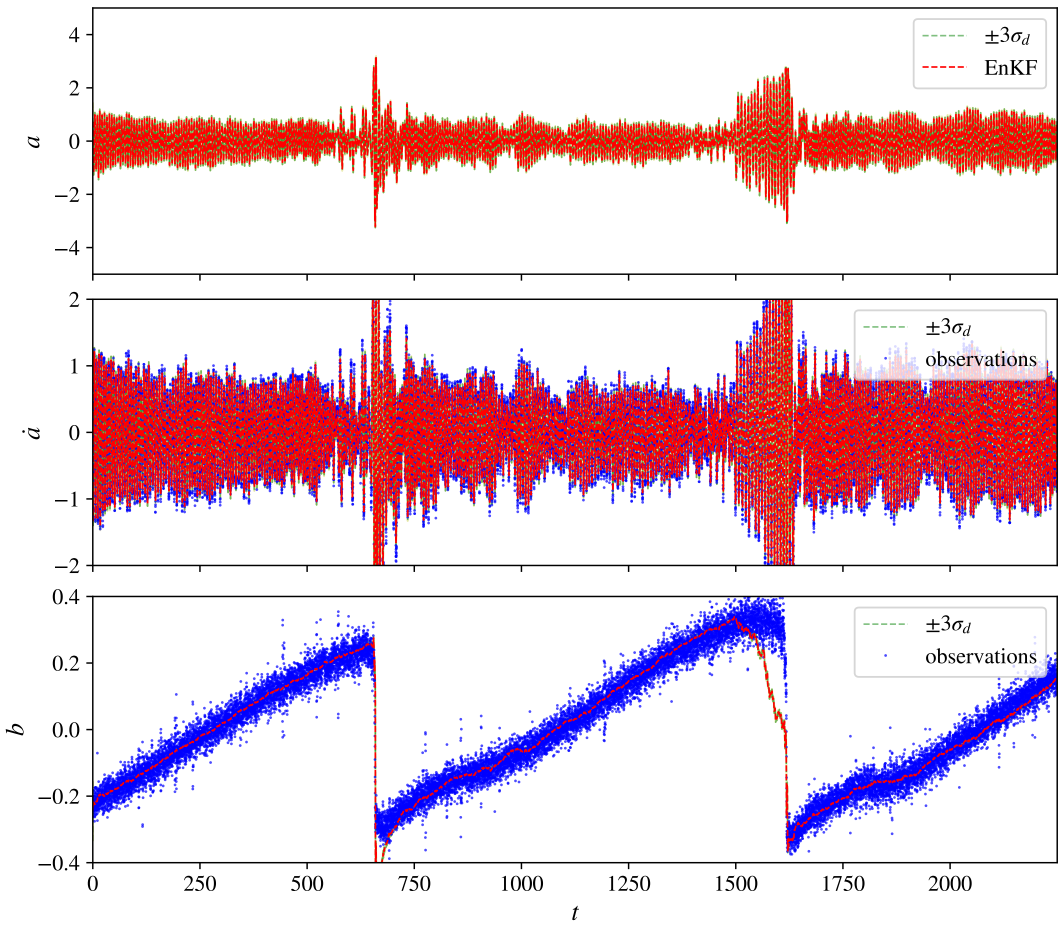


Figure 150: Reconstruction of the hidden state a and matching of observations over entire sawtoothing section of data using EnKF. The red dashed lines represent the ensemble mean and σ_d represents one standard deviation of uncertainty (posteriors following assimilation). The y scale for $\dot{a}$ is limited as the spiking in the Mirnov signal is at a large amplitude and results in the rest of the signal not being visible.

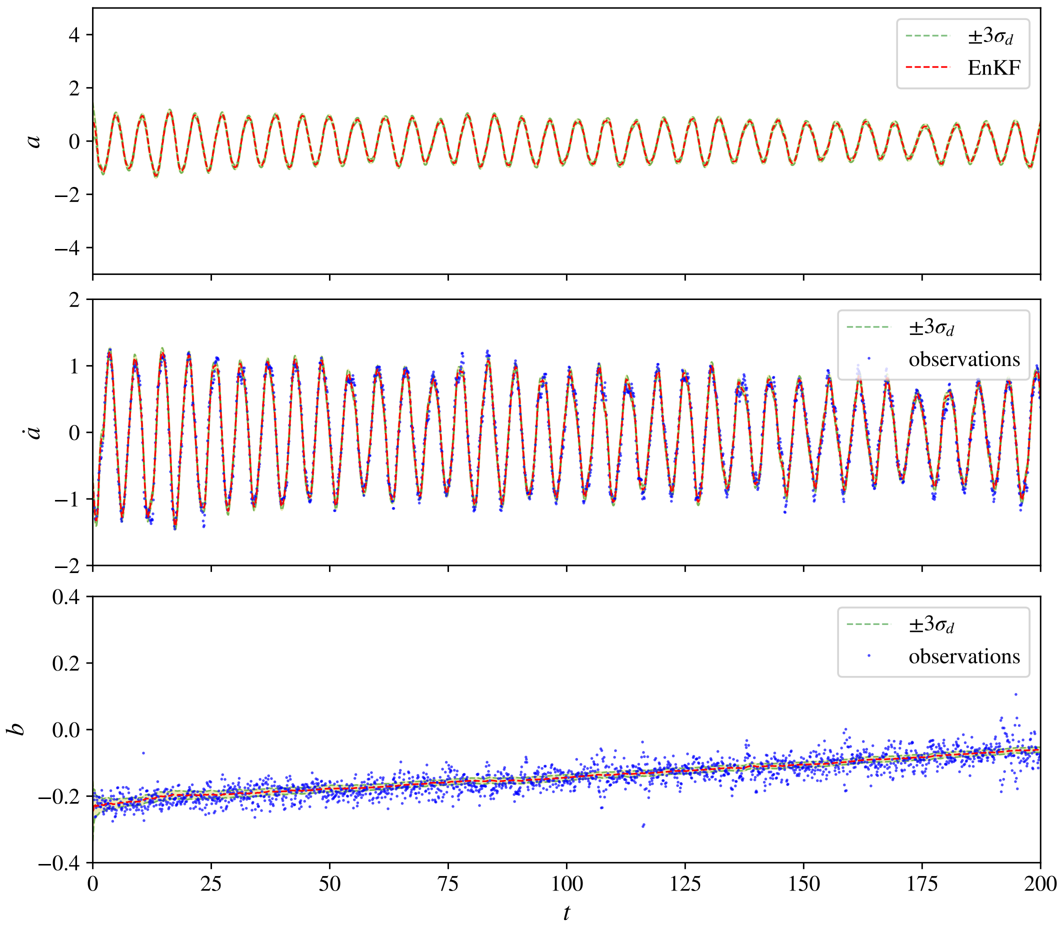


Figure 151: Reconstruction of the hidden state a and matching of observations over a short sectioning of the ramping phase during a sawtooth instability using EnKF. The red dashed lines represent the ensemble mean and σ_d represents one standard deviation of uncertainty (posteriors following assimilation).

For future work with SINDy, the results presented in §4 could be extended to the experimental data presented in Figure 150. For any application of SINDy to experimental data it would be highly advisable to implement the weak form given the degree of noise present. Observed sampling rates of $\nu \approx 60$ in the Mirnov signal further provide confidence that the weak approach could cope. Results shown in Figure 60 showed promising recovery of the Lorenz equations with similar degrees of noise and sampling rates as in the Mirnov signal. However, this is in the restrictive case where a smaller library is used and extensions to larger libraries caused significant challenges. The form of the ANAET model does include high-order nonlinearities which imply that we may need to consider large feature libraries when using SINDy. It would then be required to enforce constraints such as the symmetry constraints outlined in §2.4.2 to reduce the library complexity. Further, robustness to noise and sampling rates could be gained by considering constraints of ideal MHD that require energy conservation. For many instabilities ideal MHD plays a prominent role in their understanding [34] and could consist of a viable set of constraints for SINDy models. These types of constraints have already been explored with fluids in refs. [113, 149] and in a plasma context by ref. [150].

For successful application of SINDy to the data outlined in Figure 150, we would also need to consider the reconstruction of a by taking the integral of $\dot{a}$ (corresponding to the integral of the Mirnov signal). In principle with the sampling rates observed in modern MAST-U shots this could be possible, though introduction of a integration constant in a would result in an arbitrary translation of the model. This would have to be addressed if we wished for the outlined symmetry constraints to remain valid. We would also need to consider the assumption that a complete set of measurements are given by the Mirnov signal, the integral of the Mirnov signal and the corresponding soft X-ray measurements. One solution to this is the use of say, a false-nearest neighbours approach given in ref. [21] which seeks to estimate the embedding dimension from a single time series measurement. Consistency of the embedding dimension of the ANAET model and the experimental data would then provide evidence for (or against) the completeness of the chosen measurements. The other alternative is to rely on an embedding for reconstruction of SINDy models [122, 123]. An embedding approach attempts to reconstruct SINDy type models by only using partial observations of the system by constructing what is known as a delay matrix. Other more recent work combines an embedding approach with the extended Kalman filter for partially observed systems [173]. This is particularly interesting as it offers both noise robustness which is required in our case but also an approach for coping with partial measurements. The main issue with many of the embedding techniques is whether SINDy is now functioning as more of a “black-box” approach as the physical meaning of embedded co-ordinates becomes less clear.

Compared to EnKF, SINDy has the advantage that using a constrained polynomial library could identify which models are relevant for application with experimental data. As shown in Figure 150, there are aspects of the data which are not well represented by the underlying model in EnKF. Further, SINDy could also be used to address the question of sensible initial parameters for initialisation with EnKF. An ensembling approach as given in ref. [156] could also quantify the distributions of the model parameters and allow for a better quantification of what the initial parameter uncertainties should be when using EnKF. With the ensembling approach, the data can be bagged and several models fit over sections of the data which allows for parameter distributions to be constructed. The downsides with SINDy are, however, significant. Translation of input time series such as a by an arbitrary constant result in different models being discovered. Further, as the feature library will be at best an approximate basis for the dynamics, it is not clear that SINDy will be able to recover sparse models. The assumption of Gaussian additive noise in the regression problem will be violated in cases where the library is an approximation of the dynamics, and this produces non-sparse models. In the event SINDy does not return a sparse model, there are no clear approaches to understand why the method has not worked. Reasons such as: poor data, incomplete libraries, incorrect hyperparameter choices, excessive noise and correlated library features all result in the discovery of non-sparse models.

Other applications

Other applications could involve applying SINDy to PDE data in the context discussed in §5 where SINDy is applied to POD modes from PDE data. One of the original intentions of this project was to apply SINDy to BOUT++ data of ELMs discussed by ref. [83], but at the time of researching BOUT++ code was limited to at most the nonlinear onset of type-I ELMs. In simpler terms, this meant that only one ELM cycle was ever simulated and not repeated ELMs due to the inherent difficultly in doing so. Much of this difficult revolves around the need to resolve long quiescent periods between instabilities and fast ideal MHD timescales at the instability [131]. Even exploration of the code with completely non-physical diffusivity still did not provide alleviation in our tests and so this remains for future work. Application of SINDy to these types of problems can be performed in one of two ways: either we perform some type of modal decomposition like POD, or we attempt to fit SINDy models to carefully chosen time series measurements of simulations as in ref. [165]. In any approach, SINDy could be used to construct low-dimensional models of the PDE behaviour and if parameterised could provide information on the bifurcation structure of the PDE system. The effectiveness of models based on a modal decomposition will naturally be limited by the decomposition method itself. Referring to POD, if the singular value spectrum decreases slowly it may no longer be feasible to construct a SINDy model as many modes will have to be included.

References

1. Taylor, G. I. VIII. Stability of a viscous liquid contained between two rotating cylinders. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **223,** 289–343. eprint: `https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.1923.0008`. `https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1923.0008` (1923).
2. Chandrasekhar, S. *Hydrodynamic and Hydromagnetic Stability* (Dover Publications, Inc, 1961).
3. Kalman, R. E. & Bucy, R. S. New Results in Linear Filtering and Prediction Theory. *Journal of Basic Engineering* **83,** 95–108. ISSN: 0021-9223. eprint: `https://asmedigitalcollection.asme.org/fluidsengineering/article-pdf/83/1/95/5503549/95\_1.pdf`. `https://doi.org/10.1115/1.3658902` (Mar. 1961).
4. von Goeler, S., Stodiek, W. & Sauthoff, N. Studies of Internal Disruptions and $m = 1$ Oscillations in Tokamak Discharges with Soft—X-Ray Tecniques. *Phys. Rev. Lett.* **33,** 1201–1203. `https://link.aps.org/doi/10.1103/PhysRevLett.33.1201` (20 Nov. 1974).
5. Kadomtsev, B. B. Disruptive instability in tokamaks. *Sov. Tech. Phys. Lett. (Engl. Transl.); (United States)* **1.** `https://www.osti.gov/biblio/7147025` (May 1975).
6. Hax, A. C., Bradley, S. P. & Magnanti, T. L. *Applied Mathematical Programming* `https://web.mit.edu/15.053/www/` (Addison-Wesley, 1977).
7. Knobloch, E., Weiss, N. O. & Costa, L. N. D. Oscillatory and steady convection in a magnetic field. *Journal of Fluid Mechanics* **113,** 153–186 (1981).
8. Weiss, N. O. Convection in an imposed magnetic field. Part 1. The development of nonlinear convection. *Journal of Fluid Mechanics* **108,** 247–272 (1981).
9. Proctor, M. R. E. & Weiss, N. O. Magnetoconvection. *Reports on Progress in Physics* **45,** 1317–1379. `https://doi.org/10.1088/0034-4885/45/11/003` (Nov. 1982).
10. Wagner, F. *et al.* Regime of Improved Confinement and High Beta in Neutral-Beam-Heated Divertor Discharges of the ASDEX Tokamak. *Phys. Rev. Lett.* **49,** 1408–1412. `https://link.aps.org/doi/10.1103/PhysRevLett.49.1408` (19 Nov. 1982).
11. Arter, W. Nonlinear convection in an imposed horizontal magnetic field. *Geophysical & Astrophysical Fluid Dynamics* **25,** 259–292. eprint: `https://doi.org/10.1080/03091928308221752`. `https://doi.org/10.1080/03091928308221752` (1983).
12. Hindmarsh, A. C. ODEPACK, A Systematized Collection of ODE Solvers. *IMACS Transactions on Scientific Computation* **1,** 55–64 (1983).
13. Knobloch, E. & Weiss, N. O. Bifurcations in a model of magnetoconvection. *Physica D: Nonlinear Phenomena* **9,** 379–407. ISSN: 0167-2789. `https://www.sciencedirect.com/science/article/pii/0167278983902798` (1983).
14. Edwards, A. W. *et al.* Rapid Collapse of a Plasma Sawtooth Oscillation in the JET Tokamak. *Phys. Rev. Lett.* **57,** 210–213. `https://link.aps.org/doi/10.1103/PhysRevLett.57.210` (2 July 1986).
15. Roberts, D. E., Villiers, J. A. M. D., Fletcher, J. D., O'Mahony, J. R. & Joel, A. Major disruptions of low aspect ratio tokamak plasmas caused by thermal instability. *Nuclear Fusion* **26,** 785. `https://dx.doi.org/10.1088/0029-5515/26/6/007` (June 1986).
16. Wesson, J. *Tokamaks* (Clarendon Press - Oxford, 1987).
17. Arter, W. Phenomenological modeling of Mirnov oscillations. *Physics of Fluids - PHYS FLUIDS* **31,** 2051–2053 (July 1988).
18. Duperrex, P. *Measurement of Magnetic Fluctuations in the JET and TCA tokamak* PhD thesis (Ècole Polytechnique Fèdèrale de Lausanne, 1988).

19. Wesson, J. A. *et al.* Disruptions in JET. *Nuclear Fusion* **29,** 641. https://dx.doi.org/10.1088/0029-5515/29/4/009 (Mar. 1989).

20. Duperrex, P. A., Pochelon, A., Edwards, A. W. & Snipes, J. A. Global sawtooth instability measured by magnetic coils in the JET Tokamak. *Nuclear Fusion* **32,** 1161. https://dx.doi.org/10.1088/0029-5515/32/7/I07 (July 1992).

21. Kennel, M. B., Brown, R. & Abarbanel, H. D. I. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A* **45,** 3403–3411. https://link.aps.org/doi/10.1103/PhysRevA.45.3403 (6 Mar. 1992).

22. Biskamp, D. *Nonlinear Magnetohydrodynamics* (Cambridge University Press, 1993).

23. Chossat, P. & Iooss, G. *The Couette-Taylor Problem* (Springer-Verlag, 1994).

24. Evensen, G. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans* **99,** 10143–10162. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/94JC00572. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572 (1994).

25. Rucklidge, A. M. Chaos in magnetoconvection. *Nonlinearity* **7,** 1565. https://dx.doi.org/10.1088/0951-7715/7/6/003 (Nov. 1994).

26. Arter, W. & Edwards, D. N. Application of some novel methods of time series analysis to tokamak data. *Euroatom/ UKAEA* (1996).

27. Porcelli, F., Boucher, D. & Rosenbluth, M. N. Model for the sawtooth period and amplitude. *Plasma Physics and Controlled Fusion* **38,** 2163. https://dx.doi.org/10.1088/0741-3335/38/12/010 (Dec. 1996).

28. Zohm, H. Edge localized modes (ELMs). *Plasma Physics and Controlled Fusion* **38,** 105–128. https://doi.org/10.1088/0741-3335/38/2/001 (Feb. 1996).

29. Evensen, G. & Fabio, N. solving for the Generalized Inverse of the Lorenz Model. *Journal of Meteorological Society of Japan* **75,** 229–243 (1997).

30. Hastie, R. J. Sawtooth instability in tokamak plasma. *Astrophysics and Space Science* **256,** 177–204 (1997).

31. Burgers, G., van Leeuwen, P. J. & Evensen, G. Analysis Scheme in the Ensemble Kalman Filter. *Monthly Weather Review* **126,** 1719–1724. https://journals.ametsoc.org/view/journals/mwre/126/6/1520-0493_1998_126_1719_asitek_2.0.co_2.xml (1998).

32. Anderson, J. L. & Anderson, S. L. A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts. *Monthly Weather Review* **127,** 2741–2758. https://journals.ametsoc.org/view/journals/mwre/127/12/1520-0493_1999_127_2741_amciot_2.0.co_2.xml (1999).

33. ITER Physics Expert Group on Confinement and Transport and ITER Physics Expert Group on Confinement Modelling and Database and ITER Physics Basis Editors. Chapter 2: Plasma confinement and transport. *Nuclear Fusion* **39,** 2175. https://dx.doi.org/10.1088/0029-5515/39/12/302 (Dec. 1999).

34. ITER Physics Expert Group on Disruptions, Plasma Control, and MHD and ITER Physics Basis Editors. Chapter 3: MHD stability, operational limits and disruptions. *Nuclear Fusion* **39,** 2251. https://dx.doi.org/10.1088/0029-5515/39/12/303 (Dec. 1999).

35. Chatterjee, A. An introduction to proper orthogonal decomposition. *Curr. Sci.,* 169259–169271 (2000).

36. Gloubitsky, M., LeBlanc, V. G. & Melbourne, I. Hopf Bifurcation from Rotating Waves and Patterns in Physical Space. *Journal of Nonlinear Science* **10,** 69–101. ISSN: 0938-8974 (2000).

37. Suttrop, W. The physics of large and small edge localized modes. *Plasma Physics and Controlled Fusion* **42,** A1–A14. https://doi.org/10.1088/0741-3335/42/5a/301 (May 2000).

38. Hamill, T. M. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* **129,** 550–560. https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml (2001).

39. Simon, D. & Chia, T. L. Kalman filtering with state equality constraints. *IEEE Transactions on Aerospace and Electronic Systems* **38,** 128–136 (2002).

40. Wilson, H. R., Snyder, P. B., Huysmans, G. T. A. & Miller, R. L. Numerical studies of edge localized instabilities in tokamaks. *Physics of Plasmas* **9,** 1277–1286. eprint: https://doi.org/10.1063/1.1459058. https://doi.org/10.1063/1.1459058 (2002).

41. Boyd, T. J. M. & Sanderson, J. J. *The Physics of Plasmas* (Cambridge University Press, 2003).

42. Hazeltine, R. D. & Meiss, J. D. *Plasma Confinement* (Dover Publications, Mineola, New York, 2003).

43. Sprott, J. C. *Chaos and Time-Series Analysis* (Oxford University Press, 2003).

44. Tipping, M. E. & Faul, A. C. *Fast Marginal Likelihood Maximisation for Sparse Bayesian Models* in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (eds Bishop, C. M. & Frey, B. J.) **R4.** Reissued by PMLR on 01 April 2021. (PMLR, June 2003), 276–283. https://proceedings.mlr.press/r4/tipping03a.html.

45. Zohm, H. *et al.* MHD limits to tokamak operation and their control. *Plasma Physics and Controlled Fusion* **45,** A163. https://dx.doi.org/10.1088/0741-3335/45/12A/012 (Nov. 2003).

46. Evensen, G. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics* **54,** 539–560 (2004).

47. Julier, S. J. & Uhlmann, J. K. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE* **92,** 401–422 (2004).

48. Kantz, H. & Schreiber, T. *Nonlinear Time Series Analysis* second (Cambridge university press, 2004).

49. Tipping, M. E. in *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (eds Bousquet, O., von Luxburg, U. & Rätsch, G.) 41–62 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2004). ISBN: 978-3-540-28650-9. https://doi.org/10.1007/978-3-540-28650-9_3.

50. Zhang, F., Snyder, C. & Sun, J. Impacts of Initial Estimate and Observation Availability on Convective-Scale Data Assimilation with an Ensemble Kalman Filter. *Monthly Weather Review* **132,** 1238–1253. https://journals.ametsoc.org/view/journals/mwre/132/5/1520-0493_2004_132_1238_ioieao_2.0.co_2.xml (2004).

51. Desroziers, G., Berre, L., Chapnik, B. & Poli, P. Diagnosis of observation, background and analysis-error statistics in observation space. *Quarterly Journal of the Royal Meteorological Society* **131,** 3385–3396. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/qj.05.108. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.108 (2005).

52. Huysmans, G. T. A. ELMs: MHD instabilities at the transport barrier. *Plasma Physics and Controlled Fusion* **47,** B165–B178. https://doi.org/10.1088/0741-3335/47/12b/s13 (Nov. 2005).

53. Oyama, N. *et al.* Energy loss for grassy ELMs and effects of plasma rotation on the ELM characteristics in JT-60U. *Nuclear Fusion* **45,** 871–881. https://doi.org/10.1088/0029-5515/45/8/014 (July 2005).
54. de Blank, H. J. MHD Instabilities in Tokamaks. *Fusion Science and Technology* **49,** 118–130. eprint: https://doi.org/10.13182/FST06-A1111. https://doi.org/10.13182/FST06-A1111 (2006).
55. Evensen, G. *Data assimilation. The ensemble Kalman filter* ISBN: 9783642037108 (Jan. 2006).
56. Koslowski, H. R. Operational Limits and Limiting Instabilities in Tokamak Machines. *Fusion Science and Technology* **49,** 147–154. eprint: https://doi.org/10.13182/FST06-A1114. https://doi.org/10.13182/FST06-A1114 (2006).
57. Bongard, J. & Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **104,** 9943–9948. ISSN: 0027-8424. eprint: https://www.pnas.org/content/104/24/9943.full.pdf. https://www.pnas.org/content/104/24/9943 (2007).
58. Freidberg, J. *Plasma physics and fusion energy* (Cambridge university press, 2007).
59. Julier, S. J. & LaViola, J. J. On Kalman Filtering With Nonlinear Equality Constraints. *IEEE Transactions on Signal Processing* **55,** 2774–2784 (2007).
60. Kirk, A. *et al.* Evolution of the pedestal on MAST and the implications for ELM power loadings. *Plasma Physics and Controlled Fusion* **49,** 1259–1275. https://doi.org/10.1088/0741-3335/49/8/011 (July 2007).
61. Lang, L., Chen, W.-s., Bakshi, B. R., Goel, P. K. & Ungarala, S. Bayesian estimation via sequential Monte Carlo sampling—Constrained dynamic systems. *Automatica* **43,** 1615–1622. ISSN: 0005-1098. https://www.sciencedirect.com/science/article/pii/S0005109807001653 (2007).
62. Connor, J. W., Kirk, A. & Wilson, H. R. Edge Localised Modes (ELMs): Experiments and Theory. *AIP Conference Proceedings* **1013,** 174–190. eprint: https://aip.scitation.org/doi/pdf/10.1063/1.2939030. https://aip.scitation.org/doi/abs/10.1063/1.2939030 (2008).
63. E. J. Strait E. D. Fredrickson, J.-M. M. & Takechi, M. Chapter 2: Magnetic Diagnostics. *Fusion Science and Technology* **53,** 304–334. eprint: https://doi.org/10.13182/FST08-A1674. https://doi.org/10.13182/FST08-A1674 (2008).
64. Ingesson, L. C., Alper, B., Peterson, B. J. & Vallet, J.-C. Chapter 7: Tomography Diagnostics: Bolometry and Soft-X-Ray Detection. *Fusion Science and Technology* **53,** 528–576 (Feb. 2008).
65. Prakash, J., Patwardhan, S. C. & Shah, S. L. *Constrained state estimation using the ensemble Kalman filter* in *2008 American Control Conference* (2008), 3542–3547.
66. Wang, D. & Ruuth, S. J. Variable step-size implicit-explicit linear multistep methods for time-dependent partial differential equations. *Journal of Computational Mathematics* **26,** 838–855. ISSN: 02549409, 19917139. http://www.jstor.org/stable/43693484 (2025) (2008).
67. Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y. & Toth, Z. Ensemble Data Assimilation with the NCEP Global Forecast System. *Monthly Weather Review* **136,** 463–482. https://journals.ametsoc.org/view/journals/mwre/136/2/2007mwr2018.1.xml (2008).
68. Young, K. M. Chapter 1: Plasma Measurements: An Overview of Requirements and Status. *Fusion Science and Technology* **53,** 281–303. eprint: https://doi.org/10.13182/FST08-A1673. https://doi.org/10.13182/FST08-A1673 (2008).

69. Aanonsen, S. I., Nœvdal, G., Oliver, D. S., Reynolds, A. C. & Vallès, B. The Ensemble Kalman Filter in Reservoir Engineering—a Review. *SPE Journal* **14,** 393–412. ISSN: 1086-055X. eprint: https://onepetro.org/SJ/article-pdf/14/03/393/2554039/spe-117274-pa.pdf. https://doi.org/10.2118/117274-PA (Sept. 2009).

70. Anderson, J. Spatially and temporally varying adaptive covariance inflation for ensemble filters. *Tellus A* (Jan. 2009).

71. Arter, W. Prior Information for Nonlinear Modelling of Tokamaks. *EURATOM/UKAEA Fusion Association* (2009).

72. Arter, W. Symmetry Constraints on the Dynamics of Magnetically Confined Plasma. *Physical review letters* **102,** 195004 (June 2009).

73. Dudson, B. D., Umansky, M. V., Xu, X. Q., Snyder, P. B. & Wilson, H. R. BOUT++: A framework for parallel plasma fluid simulations. *Computer Physics Communications* **180,** 1467–1480. ISSN: 0010-4655. https://www.sciencedirect.com/science/article/pii/S0010465509001040 (2009).

74. Schmidt, M. & Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science* **324,** 81–85. ISSN: 0036-8075. eprint: https://science.sciencemag.org/content/324/5923/81.full.pdf. https://science.sciencemag.org/content/324/5923/81 (2009).

75. Wang, D., Chen, Y. & Cai, X. State and parameter estimation of hydrologic models using the constrained ensemble Kalman filter. *Water Resources Research* **45.** eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008WR007401. https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2008WR007401 (2009).

76. Witvoet, G., Westerhof, E., Steinbuch, M., Doelman, N. J. & de Baar, M. R. *Control oriented modeling and simulation of the sawtooth instability in nuclear fusion tokamak plasmas* in *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference* (2009), 1360–1366.

77. Chapman, I. T. Controlling sawtooth oscillations in tokamak plasmas. *Plasma Physics and Controlled Fusion* **53,** 013001. https://dx.doi.org/10.1088/0741-3335/53/1/013001 (Nov. 2010).

78. Ongena, J. & Oost, G. V. Energy for Future Centuries: Prospects for Fusion Power as a Future Energy Source. *Fusion Science and Technology* **57,** 3–15. eprint: https://doi.org/10.13182/FST10-A9391. https://doi.org/10.13182/FST10-A9391 (2010).

79. Yamada, M., Kulsrud, R. & Ji, H. Magnetic reconnection. *Rev. Mod. Phys.* **82,** 603–664. https://link.aps.org/doi/10.1103/RevModPhys.82.603 (1 Mar. 2010).

80. Arter, W. The equivalence between magnetoconvection and reduced magnetohydrodynamics. *Preprint CCFE-R(11)15.* https://scientific-publications.ukaea.uk/wp-content/uploads/CCFE-R1115.pdf (2011).

81. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. *Algorithms for Hyper-Parameter Optimization* in *Advances in Neural Information Processing Systems* (eds Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F. & Weinberger, K.) **24** (Curran Associates, Inc., 2011). https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

82. Chartrand, R. Numerical Differentiation of Noisy, Nonsmooth Data. *ISRN Applied Mathematics* **2011,** 164564. ISSN: xxxx-xxxx. https://doi.org/10.5402/2011/164564 (May 2011).

83. Dudson, B. D., Xu, X. Q., Umansky, M. V., Wilson, H. R. & Snyder, P. B. Simulation of edge localized modes using BOUT++. *Plasma Physics and Controlled Fusion* **53,** 054005. https://doi.org/10.1088/0741-3335/53/5/054005 (Mar. 2011).

84. Murari, A. *et al. Symmetry Based Analysis of Macroscopic Instabilities in Thermonuclear Plasmas* 2011.

85. Stewart, I. & Golubitsky, M. *Fearful Symmetry. Is God a Geometer?* (Dover Publications, 2011).

86. Xu, X. Q. *et al.* Nonlinear ELM simulations based on a nonideal peeling–ballooning model using the BOUT++ code. *Nuclear Fusion* **51,** 103040. `https://doi.org/10.1088/0029-5515/51/10/103040` (Sept. 2011).

87. Arter, W. *Blue Sky Solutions to the Magnetohydrodynamic Trigger Problem* Mar. 2012.

88. McCracken, G. & Stott, P. *Fusion, the energy of the universe* (Elsevier Academic Press, 2012).

89. Murari, A. *et al. Identifying JET instabilities with neural networks* in *2012 16th IEEE Mediterranean Electrotechnical Conference* (2012), 932–935.

90. Schoor, M. V. & Weynants, R. R. Fusion Machines. *Fusion Science and Technology* **61,** 39–45. eprint: `https://doi.org/10.13182/FST12-A13491`. `https://doi.org/10.13182/FST12-A13491` (2012).

91. Whitaker, J. S. & Hamill, T. M. Evaluating Methods to Account for System Errors in Ensemble Data Assimilation. *Monthly Weather Review* **140,** 3078–3089. `https://journals.ametsoc.org/view/journals/mwre/140/9/mwr-d-11-00276.1.xml` (2012).

92. Billings, S. A. in *Nonlinear System Identification* 17–59 (John Wiley & Sons, Ltd, 2013). ISBN: 9781118535561. eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118535561.ch2`. `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118535561.ch2`.

93. Lang, P. T. *et al.* ELM control strategies and tools: status and potential for ITER. *Nuclear Fusion* **53,** 043004. `https://doi.org/10.1088/0029-5515/53/4/043004` (Mar. 2013).

94. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D. & Cox, D. D. Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational Science and Discovery* **8,** 014008. `https://dx.doi.org/10.1088/1749-4699/8/1/014008` (July 2015).

95. Biel, W. *Status and outlook of fusion research* in. **298** (Forschungszentrum Jülich GmbH Zentralbibliothek, Verlag, Jülich, Aug. 24, 2015), 432–441. `https://juser.fz-juelich.de/record/283670`.

96. Wilson, H. R. *Edge localized modes in tokamaks* Aug. 2015. `http://hdl.handle.net/2128/10101`.

97. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* **113,** 3932–3937. ISSN: 0027-8424. eprint: `https://www.pnas.org/content/113/15/3932.full.pdf`. `https://www.pnas.org/content/113/15/3932` (2016).

98. Kutz, J., Brunton, S., Brunton, B. & Proctor, J. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems* ISBN: 978-1-611974-49-2 (Nov. 2016).

99. Mangan, N. M., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Inferring Biological Networks by Sparse Identification of Nonlinear Dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2,** 52–63 (2016).

100. Matthias Katzfuss, J. R. S. & Wikle, C. K. Understanding the Ensemble Kalman Filter. *The American Statistician* **70,** 350–357. eprint: `https://doi.org/10.1080/00031305.2016.1141709`. `https://doi.org/10.1080/00031305.2016.1141709` (2016).

101. Nathan J. Kutz, S. L. B. *Dynamic Mode Decomposition* `http://dmdbook.com/` (2021).

102. Tran, G. & Ward, R. *Exact Recovery of Chaotic Systems from Highly Corrupted Data* 2016. arXiv: `1607.01067 [math.DS]`.

103. Brunton, S. L., Brunton, B. W., Proctor, J. L., Kaiser, E. & Kutz, J. N. Chaos as an intermittently forced linear system. *Nature Communications* **8.** ISSN: 2041-1723. http://dx.doi.org/10.1038/s41467-017-00030-8 (May 2017).

104. Dam, M., Brøns, M., Juul Rasmussen, J., Naulin, V. & Hesthaven, J. S. Sparse identification of a predator-prey system from simulation data of a convection model. *Physics of Plasmas* **24,** 022310. eprint: https://doi.org/10.1063/1.4977057. https://doi.org/10.1063/1.4977057 (2017).

105. Mangan, N. M., Kutz, J. N., Brunton, S. L. & Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473,** 20170009. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2017.0009. https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2017.0009 (2017).

106. Osojnik, A. & Arter, W. *Techniques for initialising simple data assimilation calculations for plasma models* tech. rep. (Industrial Focused Mathematical Modelling, University of Oxford, 2017).

107. Roth, M., Hendeby, G., Fritsche, C. & Gustafsson, F. The Ensemble Kalman filter: a signal processing perspective. *EURASIP J. Adv. Signal Process.* **56** (2017).

108. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Science Advances* **3.** eprint: https://advances.sciencemag.org/content/3/4/e1602614.full.pdf. https://advances.sciencemag.org/content/3/4/e1602614 (2017).

109. Schaeffer, H. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473,** 20160446. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2016.0446. https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2016.0446 (2017).

110. Schaeffer, H. & McCalla, S. G. Sparse model selection via integral terms. *Phys. Rev. E* **96,** 023302. https://link.aps.org/doi/10.1103/PhysRevE.96.023302 (2 Aug. 2017).

111. Arter, W. *et al. Data assimilation approach to analysing systems of ordinary differential equations* in (May 2018), 1–5.

112. Chen, F. J. *Introduction to Plasma Physics and Controlled Fusion* 3rd ed. (Springer, 2018).

113. Loiseau, J.-C. & Brunton, S. L. Constrained sparse Galerkin regression. *Journal of Fluid Mechanics* **838,** 42–67 (2018).

114. Rudy, S., Alla, A., Brunton, S. L. & Kutz, J. N. *Data-driven identification of parametric partial differential equations* 2018. arXiv: 1806.00732 [math.NA].

115. Spinicci, L., Arter, W. & Massimiliano, R. *Using Tokamak Symmetries to model Plasma Edge Instabilities* MA thesis (Facolta di Scienze e Tecnologie ‘ Corso di Laurea in Fisica, 2018).

116. Zheng, P., Askham, T., Brunton, S. L., Kutz, J. N. & Aravkin, A. Y. *A Unified Framework for Sparse Relaxed Regularized Regression: SR3* 2018. arXiv: 1807.05411 [stat.ML].

117. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. *Optuna: A Next-generation Hyperparameter Optimization Framework* in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, Anchorage, AK, USA, 2019), 2623–2631. ISBN: 9781450362016. https://doi.org/10.1145/3292500.3330701.

118. Albers, D. J., Blancquart, P.-A., Levine, M. E., Seylabi, E. E. & Stuart, A. Ensemble Kalman methods with constraints. *Inverse Problems* **35,** 095007. https://dx.doi.org/10.1088/1361-6420/ab1c09 (Aug. 2019).

119. Brunton, S. L. & Kutz, J. N. *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2019).

120. Cartis, C., Fiala, J., Marteau, B. & Roberts, L. Improving the Flexibility and Robustness of Model-based Derivative-free Optimization Solvers. *ACM Trans. Math. Softw.* **45.** ISSN: 0098-3500. `https://doi.org/10.1145/3338517` (Aug. 2019).

121. Cartis, C. & Roberts, L. A derivative-free Gauss-Newton method. *Mathematical Programming Computation.* `https://doi.org/10.1007/s12532-019-00161-7` (2019).

122. Champion, K., Lusch, B., Kutz, J. N. & Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences* **116,** 22445–22451. ISSN: 0027-8424. eprint: `https://www.pnas.org/content/116/45/22445.full.pdf`. `https://www.pnas.org/content/116/45/22445` (2019).

123. Champion, K. P., Brunton, S. L. & Kutz, J. N. Discovery of Nonlinear Multiscale Systems: Sampling Strategies and Embeddings. *SIAM Journal on Applied Dynamical Systems* **18,** 312–333. ISSN: 1536-0040. `http://dx.doi.org/10.1137/18M1188227` (Jan. 2019).

124. Gurevich, D. R., Reinbold, P. A. K. & Grigoriev, R. O. Robust and optimal sparse regression for nonlinear PDE models. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **29,** 103113. ISSN: 1054-1500. eprint: `https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/1.5120861/14622793/103113\_1\_online.pdf`. `https://doi.org/10.1063/1.5120861` (Oct. 2019).

125. Hansen, N., Akimoto, Y. & Baudis, P. *CMA-ES/pycma on Github* Zenodo. Feb. 2019. `https://doi.org/10.5281/zenodo.2559634`.

126. Li, R., Jan, N. M., Huang, B. & Prasad, V. Constrained ensemble Kalman filter based on Kullback–Leibler divergence. *Journal of Process Control* **81,** 150–161. ISSN: 0959-1524. `https://www.sciencedirect.com/science/article/pii/S0959152418301744` (2019).

127. Mangan, N. M., Askham, T., Brunton, S. L., Kutz, J. N. & Proctor, J. L. Model selection for hybrid dynamical systems via sparse regression. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **475,** 20180534. eprint: `https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.2018.0534`. `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.2018.0534` (2019).

128. Alves, E. P. & Fiuza, F. *Data-driven discovery of reduced plasma physics models from fully-kinetic simulations* 2020. arXiv: `2011.01927 [physics.plasm-ph]`.

129. Bramburger, J. J. & Kutz, J. N. Poincaré maps for multiscale physics discovery and nonlinear Floquet theory. *Physica D: Nonlinear Phenomena* **408,** 132479. ISSN: 0167-2789. `http://dx.doi.org/10.1016/j.physd.2020.132479` (July 2020).

130. Burns, K. J., Vasil, G. M., Oishi, J. S., Lecoanet, D. & Brown, B. P. Dedalus: A flexible framework for numerical simulations with spectral methods. *Physical Review Research* **2,** 023068. arXiv: `1905.10388 [astro-ph.IM]` (Mar. 2020).

131. Cathey, A. *et al.* Non-linear extended MHD simulations of type-I edge localised mode cycles in ASDEX Upgrade and their underlying triggering mechanism. *Nuclear Fusion* **60,** 124007. `https://doi.org/10.1088/1741-4326/abbc87` (Nov. 2020).

132. de Silva, B. M. *et al. PySINDy: A Python package for the Sparse Identification of Nonlinear Dynamics from Data* 2020. arXiv: `2004.08424 [math.DS]`.

133. Deisenroth, M. P., Faisal, A. A. & Ong, C. S. *Mathematics for machine learning* (Cambridge university press, 2020).

134. Doan, N. A. K., Polifke, W. & Magri, L. Physics-informed echo state networks. *Journal of Computational Science* **47,** 101237. ISSN: 1877-7503. `https://www.sciencedirect.com/science/article/pii/S1877750320305408` (2020).

135. Jackson, A. R., Jacobsen, A. S., McClements, K. G., Michael, C. A. & Cecconello, M. Diagnosing fast ion redistribution due to sawtooth instabilities using fast ion deuterium-α spectroscopy in the mega amp spherical tokamak. *Nuclear Fusion* **60,** 126035. https://dx.doi.org/10.1088/1741-4326/abb619 (Oct. 2020).

136. Jardin, S. C., Krebs, I. & Ferraro, N. A new explanation of the sawtooth phenomena in tokamaks. *Physics of Plasmas* **27,** 032509. ISSN: 1070-664X. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/1.5140968/19765501/032509_1_online.pdf. https://doi.org/10.1063/1.5140968 (Mar. 2020).

137. Labbe, R. *Kalman and Bayesian Filters in Python* https://github.com/rlabbe/Kalman-and-Bayesian-Filters-in-Python.git. 2020.

138. Loiseau, J.-C. Data-driven modeling of the chaotic thermal convection in an annular thermosyphon. *Theoretical and Computational Fluid Dynamics* **34,** 339–365. ISSN: 1432-2250. http://dx.doi.org/10.1007/s00162-020-00536-w (July 2020).

139. Pamela, S. J. P. *et al.* Extended full-MHD simulation of non-linear instabilities in tokamak plasmas. *Physics of Plasmas* **27,** 102510. ISSN: 1070-664X. eprint: https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/5.0018208/18217213/102510_1_5.0018208.pdf. https://doi.org/10.1063/5.0018208 (Oct. 2020).

140. Raanes, P. *DAPPER* https://github.com/nansencenter/DAPPER.git. 2020.

141. Raanes, P. *Intro to data assimilation (DA) and EnKF* https://github.com/nansencenter/DA-tutorials.git. 2020.

142. Reinbold, P. A. K., Gurevich, D. R. & Grigoriev, R. O. Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Phys. Rev. E* **101,** 010203. https://link.aps.org/doi/10.1103/PhysRevE.101.010203 (1 Jan. 2020).

143. Ritchie, H. & Rosado, P. Energy Mix. *Our World in Data.* https://ourworldindata.org/energy-mix (2020).

144. Ritchie, H. & Roser, M. CO2 emissions. *Our World in Data.* https://ourworldindata.org/co2-emissions (2020).

145. Van Breugel, F., Kutz, J. N. & Brunton, B. W. Numerical Differentiation of Noisy Data: A Unifying Multi-Objective Optimization Framework. *IEEE Access* **8,** 196865–196877 (2020).

146. Virtanen, P. *et al.* SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **17,** 261–272 (2020).

147. Elfring, J., Torta, E. & van de Molengraft, R. Particle Filters: A Hands-On Tutorial. *Sensors,* 438 (2021).

148. Gaun, Y., Brunton, S. L. & Novosselov, I. Sparse nonlinear models of chaotic electroconvection. *R. Soc. Open Sci.* **8,** 202367 (2021).

149. Kaptanoglu, A. A., Callaham, J. L., Aravkin, A., Hansen, C. J. & Brunton, S. L. Promoting global stability in data-driven models of quadratic nonlinear dynamics. *Phys. Rev. Fluids* **6,** 094401. https://link.aps.org/doi/10.1103/PhysRevFluids.6.094401 (9 Sept. 2021).

150. Kaptanoglu, A. A., Morgan, K. D., Hansen, C. J. & Brunton, S. L. *Physics-constrained, low-dimensional models for MHD: First-principles and data-driven approaches* 2021. arXiv: 2004.10389 [physics.comp-ph].

151. McClements, K. G., Young, J., Garzotti, L., Jones, O. M. & Michael, C. A. Abel inversion of soft x-ray fluctuations associated with fast particle-driven fishbone instabilities in MAST plasmas. *Plasma Res. Express* **3,** 034001 (2021).

152. Messenger, D. A. & Bortz, D. M. Weak SINDy for partial differential equations. *Journal of Computational Physics* **443,** 110525. ISSN: 0021-9991. https://www.sciencedirect.com/science/article/pii/S0021999121004204 (2021).

153. Abramovic, I., Alves, E. P. & Greenwald, M. Data-driven model discovery for plasma turbulence modelling. *Journal of Plasma Physics* **88,** 895880604 (2022).

154. Callaham, J. L., Brunton, S. L. & Loiseau, J.-C. On the role of nonlinear correlations in reduced-order modelling. *Journal of Fluid Mechanics* **938,** A1 (2022).

155. Evensen, G., Vossepoel, F. C. & van Leeuwen, P. J. *Data Assimilation Fundamentals* 1st ed. (Springer Charm, 2022).

156. Fasel, U., Kutz, J. N., Brunton, B. W. & Brunton, S. L. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proc. R. Soc. A* **478,** 20210904 (2022).

157. Hirsh, S. M., Barajas-Solano, D. A. & Kutz, J. N. Sparsifying priors for Bayesian uncertainty quantification in model discovery. *R. Soc. Open Sci.,* 921182 (2022).

158. Samoylov, O., Zohm, H. & Lesch, H. *Magnetic reconnection during sawtooth instability in ASDEX upgrade* PhD thesis (Fakulät für Physik der Ludwig–Maximilians–Universität München, 2022).

159. Bakarji, J., Champion, K., Kutz, J. N. & Brunton, S. L. Discovering governing equations from partial measurements with deep delay autoencoders. *Proc. R. Soc.,* 47920230422 (2023).

160. Bertimas, D. & Gurnee, W. Learning sparse nonlinear dynamics via mixed-integer optimization. *Nonlinear Dynamics* **111,** 6585–6604 (2023).

161. Bocquet, M. & Farchi, A. *Introduction to the principles and methods of data assimilation in the geoscience* 2023.

162. Guillas, S. *Advanced Quantification of Uncertainties In Fusion modelling at the Exascale with model order Reduction (AQUIFER)* 2023.

163. Kaptanoglu, A. A., Zhang, L., Nicolaou, Z. G., Urban, F. & Brunton, S. L. Benchmarking sparse system identification with low-dimensional chaos. *Nonlinear Dynamics* **111,** 13143–13164 (2023).

164. Kaptanoglu, A. A., Hansen, C., Lore, J. D., Landreman, M. & Brunton, S. L. Sparse regression for plasma physics. *Physics of Plasmas* **30,** 033906. ISSN: 1070-664X. eprint: `https://pubs.aip.org/aip/pop/article-pdf/doi/10.1063/5.0139039/16797711/033906\_1\_online.pdf`. `https://doi.org/10.1063/5.0139039` (Mar. 2023).

165. Lore, J. D. *et al.* Time-dependent SOLPS-ITER simulations of the tokamak plasma boundary for model predictive control using SINDy. *Nuclear Fusion* **63,** 046015. `https://dx.doi.org/10.1088/1741-4326/acbe0e` (Mar. 2023).

166. Ryan, D. A. *et al.* Initial progress of the magnetic diagnostics of the MAST-U tokamak. *Review of Scientific Instruments* **94,** 073501. ISSN: 0034-6748. eprint: `https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/5.0156334/18024603/073501\_1\_5.0156334.pdf`. `https://doi.org/10.1063/5.0156334` (July 2023).

167. Tan, E. *et al.* Selecting embedding delays: An overview of embedding techniques and a new method using persistent homology. *Chaos: An Interdisciplinary Journal of Nonlinear Science* **33,** 032101. ISSN: 1054-1500. eprint: `https://pubs.aip.org/aip/cha/article-pdf/doi/10.1063/5.0137223/16786719/032101\_1\_online.pdf`. `https://doi.org/10.1063/5.0137223` (Mar. 2023).

168. Watanabe, S. *Tree-Structured Parzen Estimator: Understanding Its Algorithm Components and Their Roles for Better Empirical Performance* in *arXiv:2304.11127* (2023).

169. Wogan, N. *numbalsoda* `https://github.com/Nicholaswogan/numbalsoda`. 2023.

170. Fung, L., Fasel, U. & Juniper, M. P. *Rapid Bayesian identification of sparse nonlinear dynamics from scarce and noisy data* 2024. arXiv: `2402.15357 [stat.ME]`. `https://arxiv.org/abs/2402.15357`.

171. Gurevich, D. R., Golden, M. R., Reinbold, P. A. K. & Grigoriev, R. O. Learning fluid physics from highly turbulent data using sparse physics-informed discovery of empirical relations (SPIDER). *Journal of Fluid Mechanics* **996,** A25 (2024).

172. Gurobi Optimization, LLC. *Gurobi Optimizer Reference Manual* 2024. https://www.gurobi.com.

173. Rosafalco, L., Conti, P., Manzoni, A., Mariani, S. & Frangi, A. EKF–SINDy: Empowering the extended Kalman filter with sparse identification of nonlinear dynamics. *Computer Methods in Applied Mechanics and Engineering* **431,** 117264. ISSN: 0045-7825. https://www.sciencedirect.com/science/article/pii/S0045782524005206 (2024).

174. Steward, B. A., Cecconello, M., Bowman, C. & Team, M.-U. Reconstruction of soft x-ray emission in MAST Upgrade. *Review of Scientific Instruments* **95,** 123508. ISSN: 0034-6748. eprint: https://pubs.aip.org/aip/rsi/article-pdf/doi/10.1063/5.0219168/20293075/123508_1_5.0219168.pdf. https://doi.org/10.1063/5.0219168 (Dec. 2024).

175. Ugolini, A. R., Breschi, V., Manzoni, A. & Tanelli, M. SINDy vs Hard Nonlinearities and Hidden Dynamics: a Benchmarking Study. *IFAC-PapersOnLine* **58.** 20th IFAC Symposium on System Identification SYSID 2024, 49–54. ISSN: 2405-8963. https://www.sciencedirect.com/science/article/pii/S2405896324012837 (2024).

176. United Nations, Department of Social and Economic Affairs, Population Division. *World Population Prospects 2024: Ten Key Messages* https://population.un.org/wpp/assets/Files/WPP2024_Summary-of-Results.pdf.

177. Bevacqua, E., Schleussner, C. F. & Zscheischler, J. A year above 1.5C signals that Earth is most probably within the 20-year period that will reach the Paris Agreement limit. *Nature Climate Change* **15,** 262265 (3 2025).

178. Cannon, A. J. Twelve months at 1.5C signals earlier than expected breach of Paris Agreement threshold. *Nature Climate Change* **15,** 266269 (3 2025).

179. Arter, W. *Spiking without and with blue-sky catastrophe* To appear in: International Journal of Bifurcation and Chaos.

180. *The road to fusion energy* https://euro-fusion.org/eurofusion/roadmap/. accessed: 13/03/2025.

181. World Nuclear Association. *Plans For New Reactors Worldwide* https://world-nuclear.org/information-library/current-and-future-generation/plans-for-new-reactors-worldwide.