DATA-DRIVEN DISCRETE SPATIO-TEMPORAL MODELS: PROBLEMS, METHODS AND AN ARCTIC SEA ICE APPLICATION

JANA DE WILJES

Inauguraldissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften

vorgelegt beim Fachbereich Mathematik und Informatik
der Freien Universität Berlin
von Diplom-Mathematikerin Jana de Wiljes

Berlin, 2014

GUTACHTER:
Prof. Dr. Illia Horenko
Universita della Svizzera Italiana
Institute of Computational Science
Faculty of Informatics
Via Giuseppe Buffi 13
6900 Lugano, Schweiz

Prof. Dr.-Ing. Rupert Klein
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Institut für Mathematik
Arnimallee 2-6
14195 Berlin, Deutschland

TAG DER DISPUTATION: 01.12.2014

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

J. de Wiljes, A. Majda, and I. Horenko. An Adaptive Markov Chain Monte Carlo Approach to Time Series Clustering of Processes with Regime Transition Behavior. *Multiscale Modeling and Simulation,* 11(2):415-441, 2013.

J. de Wiljes, L. Putzig, and I. Horenko. Discrete nonhomogeneous and nonstationary logistic and Markov regression models for spatiotemporal data with unresolved external influences. *Communications in Applied Mathematics and Computational Science,* 9(1):1-46, 2014

CONTENTS

INTRODUCTION

Many natural phenomena are governed by forces on multiple spatial and temporal scales. Yet, it is often not a computationally feasible option to describe the intrinsically multiscale interactions via a deterministic model. Consequently, there is a need to go beyond purely deterministic modeling and to use stochastic processes to describe the unresolved scales of a system [60, 61, 62].

STOCHASTIC PROCESSES Stochastic processes are probabilistic extensions of deterministic processes, such as characterized by differential equations, and are defined as families of random variables indexed by totally ordered sets (usually associated with time).

The first mathematical descriptions of stochastic processes have been postulated around the turn of the 19th century in the context of examining *Brownian motion* and are attributed to Thorvald N. Thiele and Louis Bachelier. Brownian motion is the effect of random movements of particles in a substance (e.g., a gas or a fluid) caused by the surrounding molecules. Later contributions by Albert Einstein and Marian Smoluchowski [34] led to a wider recognition of the phenomenon in the physics community, laying the groundwork for *stochastic differential equations* (SDEs). The introduction of SDEs gave the opportunity to model time-dependent dynamics that have a deterministic as well as a stochastic component (e.g., white noise). From a mathematical point of view, the essential breakthrough was achieved by Kiyoshi Itō who formally developed the theory of SDEs by presenting an ansatz to integrate stochastic processes [59, 88].

Today, stochastic processes are used for a diverse spectrum of applications and are particularly relevant in the field of finance [10, 82]. Traditionally, stochastic processes are categorized according to the cardinality of their index sets and the cardinality of the corresponding state spaces. In other words, the processes are divided into the classes of continuous or discrete processes (e.g., discrete time) according to their index sets. Further, one distinguishes between processes with countable (also referred to as discrete) or uncountable (also referred to as continuous) state spaces. An example of a continuous time stochastic process is the Wiener process, which is a

mathematical model for Brownian motion. The Wiener process is particularly relevant as it is commonly used as an important component of the Itō integral [10, 82]. Typical examples of processes with a discrete time and a continuous state space are the autoregressive models [14], which are commonly employed in various application areas such as medicine [85] and volcanology [70]. In this thesis, the focus lays on stochastic processes that have a discrete state space. The paradigm of processes with a discrete state space are the Markov processes (also referred to as Markov chains for discrete time) [15, 109].

MARKOV PROCESSES The probability of a Markov process to be in a certain state only depends on the time-wise previous state. This dependency is often referred to as the Markov property. Thus, a Markov process does not have a long term memory. In fact, the Bernoulli scheme (also referred to as *Bernoulli process* for a binary state space), a memory-less model for time discrete stochastic processes, is a special case of the class of Markov processes. The stochastic description of a Markov process is given by its transition matrix, containing the probabilities to go from one state to another. As the Markov property offers a realistic description of real life systems, Markov processes are used in many applicational areas to characterize dynamics of interest, e.g., to model climate phenomena [21].

Yet, the standard Markov model does not allow to incorporate external influences that drive the considered system. A modeling approach, addressing this issue, has been proposed by Illia Horenko who suggested a model ansatz that incorporates available influences [53, 54] and is applicable to identify time discrete Markov processes with a finite state space. More specifically, the exterior quantities are taken into account by assuming that the transition matrix can be expressed by a linear combination of unknown matrices and observable quantities. Then the unknown model matrices can be identified by means of an available time-series via parametrization tools such as the FEM-BV clustering approach [52, 84].

FEM-BV CLUSTERING ANSATZ It is a common practice in the context of data-based analysis to assume stationarity of the model parameters or of the underlying probabilistic models. Yet, such a priori assumptions might render a description of the considered dynamical system unrealistic. This

problem has been addressed via the recently introduced FEM clustering framework developed by Illia Horenko [52, 84][1].

The conceptual idea is to describe a system of interest with a finite number of local stationary regimes, which can also be interpreted as clusters, and an associated explicitly time-dependent weighting process. These clusters and the corresponding hidden affiliations are determined on the basis of the given data via variational minimization of an averaged clustering functional. In general, this method has been shown to be a promising ansatz in the context of data-based model discrimination and is superior to many standard parametrization tools (e.g., k-means clustering, fuzzy clustering, hidden Markov model, artificial neural networks, and support vector machines [11, 30, 31, 52, 84]) for certain dynamical systems.

In fact, different variants (e.g., other model assumptions or models with parameters purely dependent on time) of this data-based analysis approach [84] have already been successfully applied in various application areas, for example, for the detection of hidden transitions between the stock market phases using mean daily stock return data [52] and for data-driven statistical modeling of the modes of low frequency variability of simulated southern ocean dynamics [86, 87]. In the context of inferring an appropriate description of a Markov process governed by external factors, the FEM-BV framework has shown particularly promising results, e.g, for realistic cloud modeling [53] or for the identification of voter behavior computed on the basis of weekly voter polls [54].

STANDARD APPROACHES As there are many problems that require to simulate the behavior of stochastic discrete time processes with a finite space, a wide range of standardly employed parametrization tools exists. One of the most commonly used data-based analysis tools is the *logistic model*, e.g., in applicational areas such as finance [49] and sociology [76]. It belongs to the family of generalized linear models [33, 41] and can also be derived from discrete choice models[2] [81]. The underlying principle of the model is to assume that the state of the process is associated with a utility function dependent on external factors, unknown model parameters, and noise. Further assumptions concerning the involved error processes allow to determine that the corresponding cumulative distribution function is a

1 The abbreviation FEM refers to the fact that the employed numerical technique can be linked to the fundamental idea of the finite element method (FEM).

2 The derivation by Daniel McFadden was honored with a Nobel Prize in 2000 [81].

specific logistic function dependent on the unknown model parameters and available influences.

An alternative classification tool, extensively used in numerous disciplines, is the family of *support vector machines* [27, 102]. The modeling strategy rests upon the idea of geometrical separation of the training data via an appropriately placed hyperplane that divides the considered vector space into two segments. New samples can then easily be affiliated with one of the classes. Note that support vector machines belong to the class of non-dynamical pattern recognition techniques meaning that any existing time affiliation of the available data is neglected for the determination of the hyperplane.

Artificial neural networks represent another commonly employed modeling option used to understand a discrete dynamical system with a finite state space [9, 11, 55, 69]. The architecture of an artificial neural network is motivated by the natural design of a biological synapse. An artificial neural network is not specifically fixed to have a particular structure but usually consists of (mostly *hidden*) layers of artificial neurons which are connected to each other. Each neuron is associated with weights affiliated with the input data, an activation function, and a bias. In general, the architecture of a network can be arbitrarily complex. Yet, it has been shown that two layers are already sufficient to approximate an arbitrary non-linear function [55, 68]. Consequently, artificial neural networks are successfully employed to solve relevant problems such as the diagnosis of certain cancers in the human body [39, 95].

MISSING DATA A central challenge for statistical analyses and prediction methods is the intrinsically multiscale and multiphysical nature of many natural phenomena. One of the significant manifestations of this issue is that such approaches are confronted with the problem of missing information from unresolved or unmeasured scales. Essentially, in most realistic applications not all relevant quantities are directly accessible and available in form of observations. Unfortunately, standard data-based analysis techniques often lack the option to take these missing factors into account, leading to biased and distorted results when confronted with this particular problem. As recently demonstrated by Illia Horenko in the context of modeling discrete processes, such systematically missing or implicit information can be taken into account via a non-stationary model [53, 54]. The conceptual idea is that the joint impact of all missing influences is reflected in an explicit dependency of the model parameters on time. More specifically, a Markov

model with an explicitly time-dependent transition matrix is considered. In general, the framework can be applied for the identification of discrete time processes with a finite state space driven by external quantities and given in form of ensemble observations (i.e., measurements of the relative frequency of the considered process to be in a certain state).

APPLICATION The phenomenon of accelerated melting processes in the arctic region has become the representative indicator of impending implications of the current climate change. Due to the undeniable thread posed by the melting arctic polar cap, a considerable amount of research has been conducted concentrating on producing realistic simulations of the arctic sea ice decline. In the context of describing climate phenomena, the modeling approaches, i.e., global climate models, are typically purely *deterministic*, i.e., a deterministic system that can be described via physical laws is assumed. Yet, most climate model projections of the future sea ice extent are too conservative, leading to inconsistencies with recent observations of rapid ice loss [78, 89, 108].

Additionally, phenomena such as the stagnation of the temperature rise, i.e, the rate at which the surface air temperature increases has slowed (also referred to as the *temperature hiatus* 2000-2013), raise questions which currently can not all be answered satisfactorily [36, 67]. Further, the existence of extreme negative trends concerning the sea ice extent in certain years, not explainable by natural variability alone, has been noted in [63]. In particular, the frequency of major negative trends that can not entirely be attributed to direct natural causes has increased. From a statistical point of view, such events of extreme ice decline are now expected to occur in intervals of 2 to 8 years (e.g., see ice loss in 2007 and 2012). In order to access these negative trends of sea ice related to the considerable retreat, a *stochastic* approach is suggested in [63, 118].

As elements of the climate system (e.g., ocean-to-atmosphere fluxes, surface albedo, ocean buoyancy or polar bear persistence) are directly effected by a declining arctic sea ice concentration [5, 18, 100, 104], the arctic polar cap has been closely monitored in the past decades, resulting in a vast collection of available observations. This rapid increase of collected measurements in the context of arctic sea ice variability and the growing quality of the simulation data products in recent years suggest employment of advanced data-driven approaches to gain a deeper understanding of the underlying processes. For instance, standard Markov models have been employed to predict short-term climate changes in the antarctic [21], quantile regression has

been used to determine trends in the sea ice extent in the arctic and antarctic [103, 111], and extrapolation of sea ice volume data is one approach used for the prediction of future ice concentration values [78, 89, 99]. Summarizing, parametrization of the discrete component of the arctic sea ice dynamics, based on the pure data product, might provide an unbiased (with respect to the underlying physics) insight and is thus a worthwhile attempt to gain information. On a microscopic level the corresponding dynamics have a discrete nature and thus fit the requirements for modeling via FEM-BV clustering with an a priori assumed Markov model. Yet, the process underlying the sea ice dynamics is not only evolving in time but also in space. The urge to describe natural phenomena that are dependent on time as well as on location demands that existing advanced modeling tools designed for purely time-dependent processes are extended for spatio-temporal dynamics.

OBJECTIVE In this thesis, the existing Markov regression framework is extended for modeling of discrete stochastic processes with an additional spatial component. In that context, the general problem of modeling discrete time and discrete location stochastic processes with a finite state space is contemplated.

Analogous to the purely time-dependent approach, the underlying process is assumed to have the Markov property (with respect to the time component) for the spatial enhancement. In particular, the issue of finding an adequate data-based description of the considered spatio-temporal process in the absence of relevant information is addressed. In purely time-dependent cases, unresolved governing quantities lead to a non-stationary model structure. In this thesis, it is shown that time as well as location-dependent processes that are driven by unavailable influences can be adequately described via non-stationary, non-homogenous models. A numerical approach to treat these new structural properties of the model is proposed and implemented.

In general, the inverse problem formulation corresponding to the FEM-BV clustering framework is not convex. Consequently, the methodology does not necessarily allow to compute global solutions to the problem. A Markov chain Monte Carlo clustering approach that addresses this issue is proposed. Additionally, this alternative optimization option reduces the run time for certain high dimensional problems considerably [30]. The corresponding algorithm and its implementation are given and explained.

Further, the theoretically verified abilities of the proposed non-stationary, non-homogenous Markov regression are also experimentally confirmed for an artificial test system. In particular, the characteristic property to recognize

influences that are not directly accessible is experimentally verified. Moreover, the approximations computed via the proposed model are compared to the mentioned standard approaches (i.e., logistic regression, support vector machines, and artificial neural networks).

Furthermore, the proposed framework is used in order to gain a deeper understanding of the stochastic components of the dynamics underlying the arctic sea ice extent. More precisely, a diverse range of non-stationary, non-homogenous Markov models is fitted to the satellite observations of the arctic sea ice extent. From this family of models one that is as simple as possible in terms of complexity while having a high data-reproduction quality is selected by making use of the fundamental ideas of information theory. Then this optimal model is interpreted to gain information about recent changes in the arctic sea ice coverage.

The analysis reveals particularly strong spatial correlations and a time-wise persistency that suggests that slowly evolving external processes, e.g, ocean bound forces rather than rapidly changing quantities, play a key role in the context of sea ice variability in the analyzed period (i.e., $1989 - 2004$). Moreover, individual statistical impact values of the involved measurable external factors governing the ice dynamics, such as temperature and CO_2, are computed.

OUTLINE The remainder of this thesis is structured as follows: Firstly a brief overview of three prominent standard approaches, i.e, support vector machines, artificial neural networks, and logistic regression, that can be used for modeling of discrete stochastic processes is given in Chapter 1. Then, the nature of the considered discrete stochastic process and the corresponding available ensemble data are discussed in Chapter 2. In Chapter 3 the general inverse problem associated with finding an adequate model with respect to the data is posed. Afterwards, a numerical scheme that can be used to compute a solution is presented. Although other model assumptions are mentioned, the focus is on a Markov model, which is derived in detail. Different computational options are proposed to numerically access the problem in the context of improving the quality of the results while reducing the numerical complexity. The properties of the framework are investigated by means of artificial dynamical systems in Chapter 4. A strong focus is placed upon verifying the properties of the model in the presence of unresolved quantities. Furthermore, the proposed ansatz is compared to standard approaches with respect to data-reproduction quality and out-of-sample performance.

Finally, the considered parameterization tool is employed to characterize the process underlying the arctic sea ice variability in Chapter 5. The data, the conversion to geodetic coordinates, and the settings used to identify the underlying dynamics are discussed first. Then the inferred optimal model is interpreted with respect to the behavior of the system. Further, an out-of-sample performance validation of the model and a statistical evaluation of the influences of the employed external factors are conducted. A final summary of the key findings concludes this thesis.

ACKNOWLEDGMENTS

I received a lot of support in various forms from many people to whom I would like to express my deep gratitude.

At first I would like to thank my supervisor Prof. Dr. Illia Horenko for his immeasurable amount of support and patient endurance. I particularly appreciate that he always had time, regardless of the distance between Lugano and Berlin, to provide me with his insights, to encourage me, and to help me overcome all the obstacles in the way.

My sincere gratitude is extended to Prof. Dr.-Ing. Rupert Klein for his guidance provided throughout the research process and the many helpful comments that finally led to a publication. I would also like to extent my appreciation to Prof. Dr. Andrew J. Majda for providing valuable suggestions and ideas. Further, I would like to thank Prof. Dr.-Ing. Sebastian Reich for his support during the final months of writing my thesis.

Moreover, I would like to thank my colleagues in Lugano, in particular Olga Kaiser and Lars Putzig, for the many helpful discussions, the emotional support, and the productive and also fun times spent together in Lugano and Berlin. My gratitude is extended to Dr. Philipp Metzner who always had time to answer my questions during his time as a postdoctoral fellow in Lugano. I also want to express my appreciation to the Helmholtz-Kolleg GEOSIM and the Center of Scientific Simulations for making this research possible with their financial support. Especially, I would like to thank Dr. Karen Leever and Dr. Forough Sodoudi for helping with any occurring problems.

Also, I would like to thank everyone in the Klein AG and the other GEOSIM PhD candidates. They helped to create an inspiring working environment and allowed me to get a glimpse of other interesting research fields.

My special recognition goes out to Therese, who took care of my children countless times while I sought intellectual enlightenment.

Last but not least, I would like to thank my family for their unconditional support throughout my degree: My parents, who raised me to appreciate the beauty of mathematics from an early age and who always encouraged and helped me to achieve my goals. Anne-Meike, Gertrud and Timo for providing food, love and time. My two boys, Maximilian and Sebastian, for always managing to put a smile on my face and for teaching me to put my research into perspective when it came close to consuming me. Finally, I would not have been able to accomplish this task without my husband Jan, who inspired me, who provided constant encouragement during the entire process, and who never got tired of proofreading the manuscripts. Thank you so much for your patience and love!

1

STANDARD PARAMETRIZATION APPROACHES

Let σ be a discrete stochastic process with a finite state space $\{s_1, \ldots, s_{N_S}\}$, i.e., a collection of $\{s_1, \ldots, s_{N_S}\}$-valued random variables $\{\sigma(l) : l \in \mathbb{N}\}$ indexed by a countable totally-ordered set. There is a wide range of available frameworks feasible for the data-based (i.e., observations that can be associated with the process are given) parameterization of such processes. Three of the most prominent representatives of this family, namely support vector machines [27, 102], artificial neural networks [9, 11, 55, 69], and logistic regression [33, 41, 81], are introduced in this chapter. The conceptual ideas of the three modeling approaches and their advantages and drawbacks are presented. References to more detailed discussions and introductions are given in the text.

Whenever possible, the used notations are already based on the main terminology used throughout the remainder of this thesis. In general, numbers are denoted with a capital N combined with distinguishing sub- and superscripts.

1.1 SUPPORT VECTOR MACHINES

Typically characterized as pattern recognition techniques, one of the most commonly employed classes of data-based analysis methodologies for discrete processes with a finite state space are the support vector machines (SVMs). In the following, the basic theory of standard SVMs is discussed and the reader is referred to [27, 102] for a more comprehensive introduction. The modeling ansatz of SVMs is non-dynamical meaning that the focus lays purely on geometrical properties of the data. The key idea is to geometrically separate a set of N_L (observable) training vectors $u(l) \in \mathbb{R}^{N_E}$ into different categories with respect to the given assignments $y(l) \in \{-1, 1\}$, where $l \in \{1, \ldots, N_L\}$. During the *training phase*, a universal classification rule (that can be associated with the discrete process $\sigma(l)$) is determined and can be

used to compute the class of a new sample, i.e., a sample that is not in the set of the N_L training vectors.

LINEAR SEPARABLE DATA In the most simple case, a linear separation of the vector space into two segments is possible. Then the aim is to find a hyperplane that separates the vectors $u(l)$ with maximal margins between the points of each class (see visualization in Figure 1). In detail, the points $x \in \mathbb{R}^{N_E}$ on the hyperplane are given via the equation

$$\langle w, x \rangle + m = 0, \tag{1.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the canonical inner product in $\mathbb{R}^{N_E}$. The problem is to determine $w \in \mathbb{R}^{N_E}$ and $m \in \mathbb{R}$ so that the Euclidean norm $||w||_2$ is minimal subject to the constraints

$$y(l)\,(\langle w, u(l) \rangle + m) \geq 1 \quad \forall\, l. \tag{1.2}$$

Thus, a quadratic optimization problem needs to be solved. The vectors that

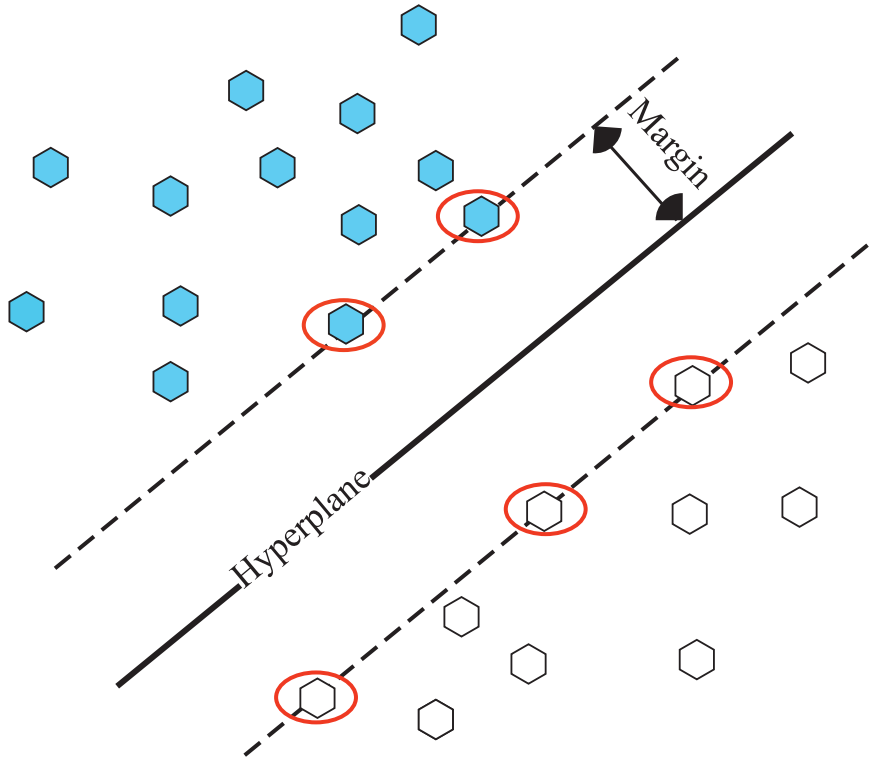

Figure 1: *A 2-dimensional example of linearly separable points in the plane is shown. The class affiliations given by $y(l)$ are visualized by blue or white coloring of the hexagons, i.e., -1 is associated with blue and 1 corresponds to white. The support vectors are marked by red ellipses.*

are on the class boundaries (see hexagons on the dashed lines in Figure 1), i.e., the vectors contained in the set

$$\left\{ u(l) \in \mathbb{R}^{N_E} \middle| \, y(l) \left(\langle w, u(l) \rangle + m \right) = 1 \ \wedge \ l \in \{1, \ldots, N_L\} \right\}, \tag{1.3}$$

are called the *support vectors*. Instead of the previously considered *primal problem* it is more convenient to solve the dual problem

$$D(\lambda) = \sum_{l=1}^{N_L} \lambda(l) - \frac{1}{2} \sum_{l=1}^{N_L} \sum_{j=1}^{N_L} y(l) y(j) \lambda(l) \lambda(j) \langle u(l), u(j) \rangle \rightarrow \max_{\lambda} \tag{1.4}$$

with the constraints

$$\lambda(l) \geq 0 \quad \forall \, l \tag{1.5}$$

and

$$\sum_{l} \lambda(l) y(l) = 0, \tag{1.6}$$

where $\lambda(l)$ are *Lagrange multipliers*. The dual optimization problem is derived using the method of Lagrange multipliers and substituting w and m with terms dependent on the unknown parameters $\lambda(l)$ [114]. In detail this is achieved by computing the first partial derivatives of the Lagrange function of the problem and by finding the corresponding extrema. It is important to note that $\lambda(l)$ corresponding to the vector $u(l)$ is equal to zero if $u(l)$ is not a support vector. Consequently, only the often much smaller subset of support vectors (see (1.3)) is required to determine the hyperplane which leads to a considerable reduction of the dimension of the problem. Note that the dual problem given in (1.4) with linear constraints (1.5) and (1.6) belongs to the class of *Quadratic programming* problems and thus can be computed with one of the many available solvers [42].

NON-SEPARABLE DATA A clear segregation of the set of vectors into two groups is not necessarily possible, e.g., due to measurement errors. Consequently, a *soft-margin* [25] that only separates most of the vectors is used, i.e., a *slack variable* $\varsigma(l) \geq 0$ is introduced and the constraint given in (1.2) is relaxed to

$$y(l)(\langle w, v(l) \rangle + m) \geq 1 - \varsigma(l) \quad \forall \, l. \tag{1.7}$$

In order to regulate the number of violations with respect to the partition, a factor $N^{\text{SVM}}_{\text{boxconstraint}}$ is used to intensify or dampen the impact of these

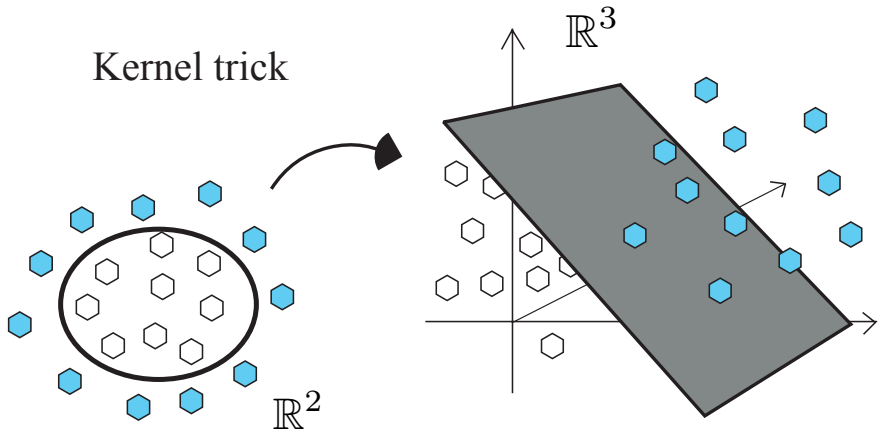

Figure 2: *Visualization of the kernel trick by means of vectors $u(l)$ with $l \in \{1, \ldots, 20\}$ in the plane that have to be mapped to a three dimensional vector space to be able to find a hyperplane that divides the considered vector into two classes. Note that the class assignments are displayed via the colors blue or white.*

errors in the corresponding optimization problem. In other words, the dual minimization problem given in (1.4) is subject to the additional constraint

$$0 \leq \lambda(l) \leq N^{\text{SVM}}_{\text{boxconstraint}} \quad \forall\, l. \tag{1.8}$$

NON-LINEAR CLASSIFICATION Unfortunately, for many applications a linear separation of a set of given vectors $u(l)$ is not very realistic. In order to circumvent this problem, the vectors $u(l) \in \mathbb{R}^{N_E}$ are linked via a function Υ to vectors in a higher dimensional vector space where a linear separation of the given points by a hyperplane is possible [1, 13]. In order to avoid the usually computer intensive (due to the required high dimensional space) calculations necessary to determine values $\Upsilon(u(l))$, a *kernel function* $\mathcal{K}$ is used that simply determines the dot product $\langle \Upsilon(u(l)), \Upsilon(u(j)) \rangle$ of the projected vectors $u(l)$, $u(j) \in \mathbb{R}^{N_E}$, which is often much easier to access numerically. This procedure is often referred to as the *kernel trick* (see visualization in Figure 2). Commonly used kernel function families are for example polynomials, i.e.,

$$\mathcal{K}^{\text{poly}}(u(l), u(j)) = \langle u(l), u(j) \rangle + 1, \tag{1.9}$$

or the Gaussian *radial basis function* (RBF), i. e.,

$$\mathcal{K}^{\text{RBF}}(u(l), u(j)) = \exp\left(\frac{||u(l) - u(j)||_2^2}{2v^2} \right) \tag{1.10}$$

for $v \in \mathbb{R}$ and the *multilayer perceptrons*, i.e.,

$$\mathcal{K}^{\mathrm{MLP}}(u(l), u(j)) = \tanh\left(r_1 \langle u(l), u(j) \rangle + r_2\right) \tag{1.11}$$

with $r_1 > 0$ and $r_2 < 0$ [79]. The multilayer perceptron variant of the SVMs is also a model class that belongs to the family of artificial neural networks (see introduction in the next section) [24]. Summarizing, the benefits of SVMs are that they provide a globally optimal classification with a relatively small computational complexity[1]. However, as the placement of the hyperplane is independent of any time (or location) affiliation of the data, it is particularly prone to provide distorted results in the presence of missing data. Another drawback is that an SVM can only be directly applied for the classification of processes with a binary state space. Yet, there are options to interpret problems with more than two classes as several binary problems [56]. Further, although the SVM often provides stable assignments, the underlying model is very abstract, in particular for a non-linear kernel function, and thus not easy to interpret. Moreover, it remains unclear to what extent the choice of a kernel function predetermines or distorts the outcomes.

1.2 ARTIFICIAL NEURAL NETWORKS

The class of artificial neural networks (ANNs) also belongs to the family of pattern reconnection techniques and ANNs are successfully employed to solve relevant problems, e.g., to diagnose several cancers [39, 95]. ANNs originally emerged in the context of biology and they are motivated by the natural design of neuronal networks in the brain or in the spinal cord. The topology of different ANNs can vary a lot but the essential building blocks are the neurons. The architecture of a single neuron is visualized in Figure 3. A considered input vector $u \in \mathbb{R}^{N_E \times 1}$ is subject to three operations going through the neuron [7]. First, the entries u_e of the vector u are weighted and summed up, i.e., a vector $\mathcal{W} \in \mathbb{R}^{1 \times N_E}$ is multiplied with input vector u. Theoretically, an alternative weighting procedure can be considered, e.g., the function computing the deviations $|u_e - w_e|$ for $e \in \{1, \ldots, N_E\}$ with w_e being the eth entry of $\mathcal{W}$ can be used [7]. Secondly, a bias $\mathfrak{b}$ is added to the sum of the weighted input values u_e. This bias is often interpreted to be an additional weight affiliated with an artificial input quantity equal to one. The sum of $\mathcal{W}$ times u and the bias forms the *net input* (see graphic in

1 A more detailed consideration of the run time of an examplary SVM can be found in Section 3.5.

Figure 3). Different variants of the currently considered net input operation are also possible but are not as common [7]. Thirdly, a *transfer function* $\Psi(\cdot)$ (also referred to as *activation function*) is applied to the net input and the resulting value is an approximation of the given target y. Commonly used

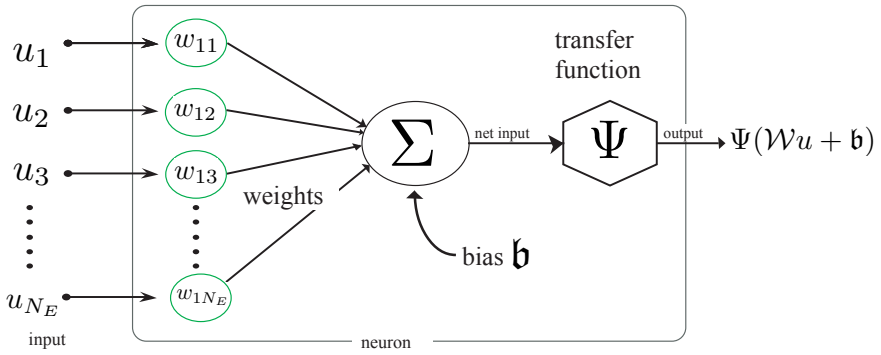

Figure 3: *A graphic interpretation of the structure of a single neuron is shown.*

transfer functions are the hyperbolic tangent function, i.e.,

$$\Psi^{\tanh}(t) = \frac{\exp(t) - \exp(-t)}{\exp(t) + \exp(-t)}, \tag{1.12}$$

and the logistic function, i.e.,

$$\Psi^{\text{sigmoid}}(t) = \frac{1}{1 + \exp(-t)}. \tag{1.13}$$

One of the advantages of these two activation functions (given in (1.12) and (1.13)) is that they are differentiable. Consequently, it is possible to employ gradient descent based training methods. Another popular option is the rectifier activation function

$$\Psi^{\text{rectifier}}(t) = \max(0, t). \tag{1.14}$$

A general network consists of several layers of the described neurons (see exemplary network in Figure 4). Specifically, the output of the neurons in the first layer (see hidden layer 1 of the exemplary network in Figure 4) is the input of the neurons in the next layer. The output layer usually has a very different structure and is designed to produce an output suiting the particular form of the given targets $y(l)$ where $l \in \{1, \ldots, N_L\}$, e.g., $y(l) \in [0, 1]$. Consequently, the output layer is treated individually, e.g., some transfer functions are typically used in the output layer but not as

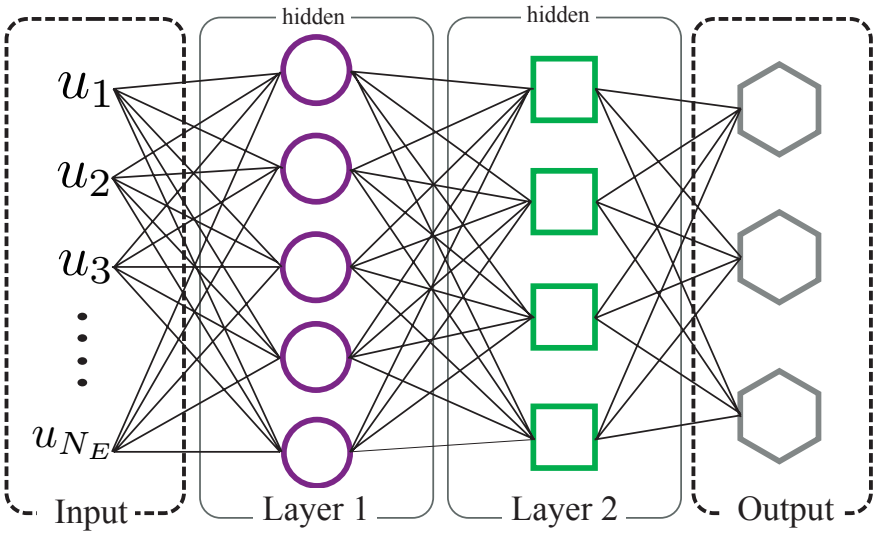

Figure 4: *A visualization of an exemplary fully connected network with two hidden layers is displayed. In detail, the first hidden layer has 5 neurons, the second hidden layers has four neurons, and the output layer consist of three neurons. The architecture of a neuron can be seen in Figure 3.*

frequently employed in the hidden layers. The layers between the input[2] and the output layer are referred to as *hidden layers* [7]. Each layer can have a different architecture, e.g., with respect to the total number of neurons or the affiliated activation functions. The exemplary network shown in Figure 4 has two hidden layers (one with five neurons and one with four neurons) and one output layer with three neurons. Summarizing, the structure of a network can be arbitrarily complex. In the training phase a network is fitted according to a set of N_L input vectors $u(l)$ and the corresponding targets $y(l)$ (affiliated with the output). In other words, an appropriate vector of weights $\mathcal{W}(l)$ and a bias $\mathfrak{b}$ have to be computed for each neuron in each layer.

One can further distinguish between different processing directions of a network, i.e., feedforward and feedbackward networks. In this thesis, the focus is on feedforward networks where the neurons of each layer are fully connected to the neurons of the neighboring layers. Further, the activation function of the neurons is chosen to be a non-linear function, e.g., the hyperbolic tangent function or the logistic function can be used. The class of networks with these properties is also referred to as multilayer perceptrons (MLPs) which have already been mentioned in the context of SVMs in the previous section [24]. As it has been shown[3] that these feedforward

2 The input is sometimes also interpreted as a layer.
3 This and related results are often referred to as the *universal approximation theorem* [68].

networks with a single hidden layer (with an arbitrary depth, i.e., number of neurons) are already sufficient to characterize most of the practically relevant functions [55], MLPs with only one hidden layer are considered for the computations in this thesis. The notation $\mathcal{N}(N^{\text{ANN}}_{\text{neurons}})$ is used for the corresponding network, where $N^{\text{ANN}}_{\text{neurons}}$ is the total number of neurons contained in the hidden layer of the considered MLP.

As there is a great variety of different networks in the class of ANNs, it is usually possible to find a network that produces good approximations of a considered systems. Yet, the selection of appropriate activation functions as well as the choice of the general architecture of the employed network is crucial and often involves a *trial and error* tuning procedure. Fortunately, it is possible to exploit that the approximation ability of rather simple (one hidden layer) MLPs has been verified to be *universal* [55]. Due to the complex internal structure of an ANN, a network is even harder to interpret than an SVM, which at least can be interpreted geometrically. Further, ANNs like the SVMs have stationary and homogenous model parameters (i.e., weights and biases). Consequently, a network confronted with systems that can not be fully described with the available data might produce distorted results.

1.3 LOGIT MODELS

In the following, a member of the class of discrete choice models, namely a logit model, is introduced. This representative of the generalized linear model family [33, 41] is a commonly employed data analysis tool in applicational areas ranging from finance [49] to sociology [76]. The derivation of the logit models via discrete choice models in the context of utility theory has first been proposed in [81]. The conceptual idea is to associate the discrete outcome of a regarded dynamical process with a cost function

$$\mathcal{C}_i[u(l), B^i] := b^i_0 + \sum_{e=1}^{N_E} b^i_e u_e(l) + \xi^i(l) \quad \text{with } i \in \{1, \dots, N_S\} \tag{1.15}$$

that depends on unknown model parameters

$$B^i = \begin{bmatrix} b^i_0 \\ \vdots \\ b^i_{N_E} \end{bmatrix} \in \mathbb{R}^{(N_E+1)\times 1}, \tag{1.16}$$

the measurable vectors $u(l) \in \mathbb{R}^{N_E \times 1}$, and on random terms $\xi^i(l)$ describing measurement errors (with $l \in \{1, \dots, N_L\}$) [75, 81]. Essentially that means that the considered discrete process is assumed to have the following form:

$$\sigma(l) = \begin{cases} s_1 & \text{if } \mathcal{C}_1[u(l), B^1] > \mathcal{C}_i[u(l), B^i] \quad \forall\, i \neq 1, \\ \vdots & \\ s_{N_S} & \text{if } \mathcal{C}_{N_S}[u(l), B^{N_S}] > \mathcal{C}_i[u(l), B^i] \quad \forall\, i \neq N_S. \end{cases} \tag{1.17}$$

Then the probability for the dynamical process $\sigma(l)$ to be in state s_i is [4]

$$\mathbb{P}\Big[\sigma(l) = s_i\Big] = \mathbb{P}\Big[\mathcal{C}_i[u(l), B^i] > \mathcal{C}_h[u(l), B^h] \;\forall\, h \neq i\Big] \tag{1.18}$$

$$= \mathbb{P}\Big[b_0^i + \sum_{e=1}^{N_E} b_e^i u_e(l) + \xi^i(l) > b_0^h + \sum_{e=1}^{N_E} b_e^h u_e(l) + \xi^h(l) \;\forall\, h \neq i\Big] \tag{1.19}$$

$$= \mathbb{P}\Big[b_0^i - b_0^h + \sum_{e=1}^{N_E} [b_e^i - b_e^h] u_e(l) + \xi^i(l) > \xi^h(l) \;\forall\, h \neq i\Big]. \tag{1.20}$$

Under the assumption that each $\xi^i(l)$ is independent and identically distributed (i.i.d.) according to the extreme value distribution (also known as Gumbel distribution), the state probabilities can be expressed as follows:

$$\mathbb{P}[\sigma(l) = s_i] = \frac{\exp\left(b_0^i + \sum_{e=1}^{N_E} b_e^i u_e(l)\right)}{\sum_{h=1}^{N_S} \exp\left(b_0^h + \sum_{e=1}^{N_E} b_e^h u_e(l)\right)} \quad \forall i. \tag{1.21}$$

This particular model, known as logit model, is one of the most prominent discrete choice models. For a detailed derivation of the state probabilities given in (1.21) from the general definition, the reader is referred to [81, 113]. Note that a variety of alternative choice models can be constructed by assuming different probability distribution functions for the random error process $\xi^1(l), \dots, \xi^{N_S}(l)$, e.g., to consider the probit models, the errors are assumed to be multivariate normal distributed.

The introduced multivariate logistic model is applicable for discrete processes with a finite state space, yet it is important to be aware of one restricting attribute of the logit model referred to as the *independence of*

4 Note that the probability of $\mathcal{C}_i[u(l), B^i] = \mathcal{C}_h[u(l), B^h]$ is assumed to be zero (see [81]).

irrelevant alternatives (IIA) property [74]. Essentially, it states that the ratio of the probabilities of any two alternatives states s_i and s_h is

$$\exp\left(b_0^i - b_0^h + \sum_{e=1}^{N_E}(b_e^i - b_e^h)u_e(l)\right). \tag{1.22}$$

As this ratio does not depend on any state other than s_i and s_h, the relative odds of a logistic model remain the same [113]. Unfortunately, this property of the model may cause interpretation problems in terms of a considered application. These difficulties are often exemplified by means of an example first formulated by McFadden in [81]: A binary state space consisting of the alternative options for an individual to take an auto or a blue bus to reach a certain destination is considered. In the example the individual chooses according to the distribution $[2/3, 1/3]$. After introducing a third alternative in form of a red bus, the often more "intuitive" probability distribution $[2/3, 1/6, 1/6]$ is not equal to the distribution derived by the logistic model, which is $[1/2, 1/4, 1/4]$. Consequently, one should avoid to employ a logistic model for applications where one of the finite number of states is a good substitute of another state. Summarizing, the considered application has to be consistent with the IIA attribute of the logistic model. Otherwise it is prudent to deploy a different, more appropriate, discrete choice model [23].

Further, it is important to mention that the intrinsic transformation, required to successfully fit a logistic model to the observed data by finding appropriate model parameters B^i describing the underlying discrete process $\sigma(l)$ with continuous regression techniques, is a map going from the closed interval $[0, 1]$ to the real numbers $(-\infty, \infty)$. This internal mapping results in stability problems on the boundaries of the logistic cumulative density function.

Yet, being aware of these drawbacks, data analysis via a logistic model ansatz is a valuable tool for various modeling scenarios of different applicational areas, in particular for dynamical systems that exhibit non-linear behavior.

2

DISCRETE SPATIO-TEMPORAL DYNAMICAL PROCESS

In this thesis, the problem of modeling spatio-temporal discrete processes[1] with a finite state space is approached. A basic introduction of the formal setting is given in the following where the focus lays, in particular, on the available information in form of observations. Further, exterior influences driving the dynamical system under consideration are discussed, where the emphasis of the discussion is on those external quantities that can not be measured. As the observations are typically on a different scale than the considered dynamics the issue of relating the spatio-temporal process to the data is contemplated as well.

For the remainder of this thesis, the discrete countable index set of the considered collection of random variables is divided and associated with time (indexed t) and locations[2] (indexed j and l), i.e., the discrete stochastic process is denoted $\sigma(t, j, l)$. Further, $\sigma(t, j, l)$ is assumed to take values in the finite set $\{s_1, \ldots, s_{N_S}\}$.

2.1 ENSEMBLE DATA AND EXTERNAL FACTORS

Measuring tools allow to take noisy snapshots of such real life dynamical processes $\sigma(t, j, l)$, i.e., $t \in \{1, \ldots, N_T\}$, $l \in \{1, \ldots, N_{\text{ens}}\}$, and $j \in \{1, \ldots, N_J\}$, where the locations are denoted $\omega(j, l)$.

Note that the considered locations $\omega(j, l)$ are associated with cells on a fine lattice (see small grid boxes in Figure 5). The index j corresponds to larger grid cells (see cells of lattice on the left in Figure 5) of the same shape, each containing N_{ens} cells of the fine grid (see cells of lattice on the right in Figure 5). The considered process, thus, assigns discrete states $s_i \in \{s_1, \ldots, s_{N_S}\}$ to each cell $\omega(j, l)$ on a microscopic grid for every t. For example, the process

1 In other words, time-discrete and location-discrete stochastic processes with a finite state space are considered.

2 A distinction is made between different spatial scales, hence two indices are used for the spatial association.

can describe the characteristic evolution of the aggregate state of water in a particular location over time t (see blue and white grid cells in Figure 5).

The considered observations provide information about the regarded system, which can be extracted with data analysis techniques. However, in most applicational areas, observations of a single realization of the dynamical process $\sigma(t, j, l)$ for fixed t, j and l are not available. Nevertheless, the relative frequency of the regarded process to be in state s_i observed on a macroscopic scale (e.g., of the size of the larger grid cells indexed $j \in \{1, \dots, N_J\}$) is often accessible via measuring. In other words, it is often possible to observe the quotient

$$\tilde{\pi}_i(t, j) = \frac{N_{s_i}(t, j)}{N_{\text{ens}}} \tag{2.1}$$

with $N_{s_i}(t, j)$ being the total number of cells $\omega(j, l)$ in state s_i, i.e.,

$$N_{s_i}(t, j) = \sum_{l=1}^{N_{\text{ens}}} \delta_{s_i}(\sigma(t, j, l)), \tag{2.2}$$

where

$$\delta_{s_i}(\sigma(t, j, l)) = \begin{cases} 1 & \text{if } \sigma(t, j, l) = s_i \text{ ,} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

is the Kronecker delta for the value s_i. This particular information on the considered dynamical process $\sigma(t, j, l)$ is referred to as *ensemble observation*. Therefore, the microscopic cells $\omega(j, l)$, corresponding to a fixed coarser grid box j, are referred to as *ensemble members*. Note that the total number of ensemble members for each $j \in \{1, \dots, N_J\}$ is N_{ens}. Due to the usually very large total number of ensemble members N_{ens}, the relative frequency (see (2.1)) is a good estimate of the probability of the process $\sigma(t, j, l)$ to be in state s_i, i.e.,

$$\pi_i(t, j) := \mathbb{P}[\sigma(t, j, l) = s_i] \approx \tilde{\pi}_i(t, j). \tag{2.4}$$

For instance, a typical observation in the context of aggregate states of water in the arctic area are satellite images providing information on the sea ice concentration. For instance, the size of one larger grid cell, indexed j, may be 25 km^2 or even larger. Consequently, the total number of microscopic locations $\omega(j, l)$ in one grid cell indexed j is indeed large. Thus, the relative frequency values obtained via satellite are good approximations of the state

probabilities $\pi_i(t,j)$ for fixed time t and coarse cell j. In the following, it is assumed that it is possible to directly observe the vector of state probabilities

$$\pi(t,j) := \begin{bmatrix} \pi_1(t,j) \\ \vdots \\ \pi_{N_S}(t,j) \end{bmatrix} \in [0,1]^{N_S \times 1}. \tag{2.5}$$

In pursuance of characterizing the dynamics of a regarded process $\sigma(t,j,l)$, it is also necessary to consider the vector of all relevant exterior influences

$$\bar{u}(t,j) \in \mathbb{R}^{N_F \times 1}. \tag{2.6}$$

Ideally, these exterior quantities $\bar{u}(t,j)$ are also observable. However, usually it is not possible to have access to all external factors having an impact on the considered dynamical process. The presence of unknown influencing quantities is one of the key problems of statistical data analysis. Therefore, it is necessary to distinguish between known and unknown quantities and to put a special focus on the unresolved factors in the parametrization procedure.

2.1.1 *Implicit external factors*

As mentioned above, the considered dynamical process $\sigma(t,j,l)$ is driven by exterior quantities comprised in the vector $\bar{u}(t,j)$. The vector of external factors $\bar{u}(t,j)$ can be divided into quantities

$$u(t,j) := \begin{bmatrix} u_1(t,j) \\ \vdots \\ u_{N_E}(t,j)) \end{bmatrix} \in \mathcal{U} \subset \mathbb{R}^{N_E \times 1} \tag{2.7}$$

that are available in form of data (also referred to as *resolved* or *explicit*) and unknown factors

$$u^{\text{unres}}(t,j) = \begin{bmatrix} u_1^{\text{unres}}(t,j) \\ \vdots \\ u_{N_I}^{\text{unres}}(t,j)) \end{bmatrix} \in \mathbb{R}^{N_I \times 1}, \tag{2.8}$$

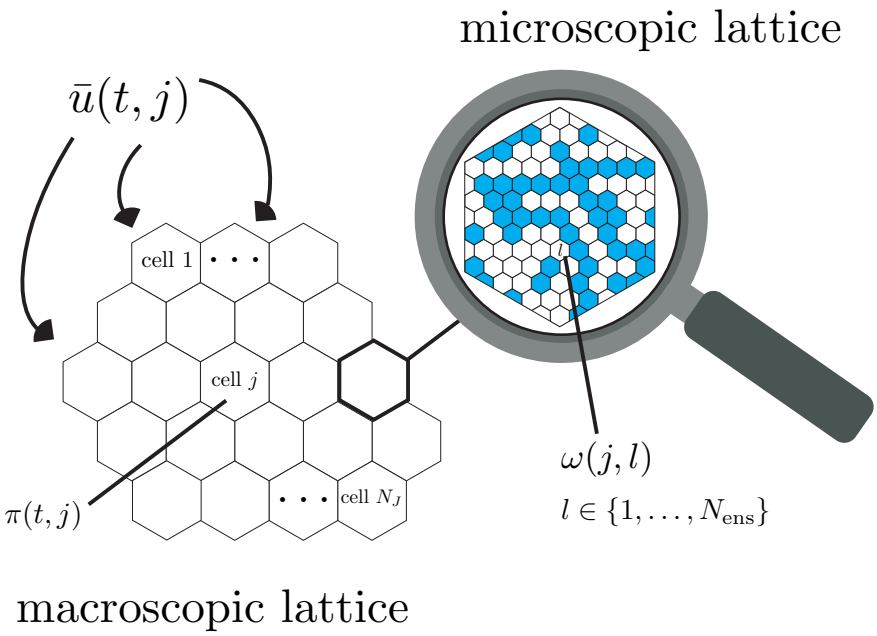

Figure 5: *A graphical interpretation of the relation between the microscopic locations $\omega(j,l)$ and the macroscopic observation $\pi(t,j)$ is presented. In order to visualize the different spatial scales, the time is fixed and a dynamical system with a binary state space is chosen, i.e., $N_S = 2$. The exemplary microscopic state assignments are shown for one macroscopic cell in the magnifying glass (see hexagonal lattice on the right). Note that the microscopic cells $\omega(j,l)$ associated with s_1 are white and the ones associated with s_2 are colored blue. The macroscopic hexagonal lattice on the left relates to the scale of the observation. In other words, for each cell indexed j it is possible to measure the corresponding $\pi(t,j)$ ensemble data. Besides showing the linkage between the scales, the images also place the focus on the additional influences on exterior quantities contained in $\bar{u}(t,j)$ which may originate from very different scales.*

which are not available as observations (also referred to as *unresolved* or *implicit*), i.e.,

$$\bar{u}(t,j) = \begin{bmatrix} u(t,j) \\ u^{\text{unres}}(t,j) \end{bmatrix} \in \mathbb{R}^{(N_E+N_I)\times 1}. \tag{2.9}$$

It is important to mention that the vector of unresolved external factors $u^{\text{unres}}(t,j)$ may consist of any quantities potentially playing a role in the dynamics of the considered process $\sigma(t,j,l)$. This includes stochastic as well as deterministic processes, in particular, processes interacting on different time-wise or location-wise scales than the ones currently regarded (i.e., $t \in \{1,\ldots,N_T\}$ and $j \in \{1,\ldots,N_J\}$). More precisely, $u^{\text{unres}}(t,j)$ may contain elements of influencing quantities associated with a microscopic scale or components that directly influence the microscopic locations $\omega(j,l)$. Further, the vector of external influences may also include any information

or quantities that only potentially play a role in the underlying dynamics. In particular, any information given via measurements can be processed in some form and then added to the vector of explicit external factors. For example, in order to identify existing spatial correlations between the neighboring cells $j \in \{1, \dots, N_J\}$, the corresponding information can be added to the vector of explicit external factors. This idea will be discussed in more detail in Section 3.4. Summarizing, there is no limitation with respect to the nature of the quantities (i.e., the entries of $\bar{u}(t, j)$) influencing the considered system on some level.

A computation of a qualitative parametrization of a process $\sigma(t, j, l)$ with standard data analysis techniques on the basis of observations $\pi(t, j)$ and $u(t, j)$ is often hampered by incomplete or missing data. The negative effect of such lack of information becomes even more pronounced if quantities with significant impact are not available as measurements, i.e., factors $u^{\text{unres}}(t, j)$ have a strong influence on the dynamics of the regarded system.

In reality, missing information $u^{\text{unres}}(t, j)$ is the status quo in most application areas due to the fact that observation tools have physical limits and the size as well as availability of data collections strongly depend on locations and on time. The arctic sea ice concentration, for example, is influenced by many quantities that are either not measurable at all or only available for certain small areas but can not be provided for the entire arctic ocean. Consequently, any technique employed to identify dynamics of interest by determining model parameters on the basis of data, missing relevant information, has to take the lack of information into account.

One way to do so is to reflect any implicit external factors in form of a joint impact in an explicit dependency on time and location. Put differently, a non-stationary, non-homogenous setting allows to incorporate dynamical influences not available in form of data. This ansatz is presented and mathematically justified in Subsection 3.1.1 for a dynamical process that is assumed to have the Markov property.

3

NON-STATIONARY NON-HOMOGENOUS AVERAGED CLUSTERING APPROACH

In this chapter, a data analysis tool based on variational minimization of a regularized clustering functional is introduced [31, 52, 84]. At first, a general inverse problem is formulated in order to parametrize a discrete dynamical system, influenced by exterior quantities, given by means of a spatio-temporal time series. Then an approach to reflect implicit external factors in an explicit dependency of the model parameters on time and locations, assuming that the considered dynamical process underlying the data has the Markov property, is proposed.

Although the considered data analysis technique is also presented in a general setting, the emphasis is specifically on the non-stationary, non-homogenous Markov model example, which will be applied to a multi-dimensional real life time series, i.e, arctic sea ice concentration data, in Chapter 5.

Further, due to the ill-posedness of the problem formulation, certain regularization steps have to be taken to address the issue. Two alternative regularization steps are presented and discussed. Then a numerical scheme for the presented optimization problem is outlined. In particular, an MCMC approach is proposed in the context of a specific Tikhonov regularized variant of the considered functional.

Then the problem of selecting an appropriate model is discussed and an information criterion is introduced. Concluding, the possibility to make statements concerning the future evolution of a dynamical process after identifying the corresponding model parameters is discussed and the associated theoretical and numerical steps are contemplated.

3.1 MODEL DISTANCE FUNCTION

Let $\{\pi(1,1), \pi(2,1), \ldots, \pi(N_T,1), \pi(2,1), \ldots, \pi(N_T,N_J)\}$ be a spatio-temporal time series with $\pi(t,j) \in [0,1]^{N_S \times 1}$ being the vector of probabilities for the underlying process $\sigma(t,j,l)$ to be in a state $s_i \in \{1, \ldots N_S\}$. The unknown model parameters, identifying the dynamics of the considered system, are denoted $\theta(\bar{u}(t,j)) \in \Omega$, where Ω is a corresponding appropriate parameter space. The relation between the parameters and the observations is assumed to be defined by a model function $f(\cdot)$, which belongs to a certain family of *direct mathematical models*, i.e.,

$$\pi(t+1,j) = f(\pi(t,j), \ldots, \pi(t-N_M,j), \theta(\bar{u}(t,j))), \tag{3.1}$$

dependent on the current and the previous observations $\pi(t,j), \ldots, \pi(t-N_M,j)$ up to a memory depth N_M and model parameters $\theta(\bar{u}(t,j))$. The model function $f(\cdot)$ can be purely deterministic or include stochastic elements, e.g.,

$$f^{\text{kmeans}}(\theta(t+1,j)) := \theta(t+1,j) + \epsilon(t+1,j), \tag{3.2}$$

where $\epsilon(t,j)$ is an i.i.d. random process with expected value zero for all t and j. Note that in this example the next state neither depends on the current one nor on the previous data values. It is also possible to consider a non-stationary and non-homogenous adaption of the proposed discrete choice models. In detail that means that the direct mathematical model is assumed to be

$$f^{\text{logit}}(B(t+1,j)) = \theta^{\text{logit}}(B(t+1,j), u(t+1,j)) + \zeta(t+1,j), \tag{3.3}$$

where the model parameter is defined as follows:

$$\theta^{\text{logit}}(B(t+1,j), u(t+1,j)) = \begin{bmatrix} \mathbb{P}[\sigma(t+1,j,l) = s_1] \\ \vdots \\ \mathbb{P}[\sigma(t+1,j,l) = s_{N_S}] \end{bmatrix} \in \mathbb{R}^{N_S \times 1} \tag{3.4}$$

with space- and time-dependent model parameters

$$B(t,j) = \left[B^1(t,j), \ldots, B^{N_S}(t,j)\right] \in \mathbb{R}^{(N_E+1)\times N_S}. \tag{3.5}$$

The general problem of finding model parameters $\theta(\bar{u}(t,j))$ that explain the observations $\pi(t,j)$ "best" with respect to a priorly chosen model function

$f(\cdot)$ is referred to as *inverse problem*. To be able to measure the *fitness* of a set of model parameters $\theta(\bar{u}(t,j))$ with respect to the data $\pi(t,j)$, a *model distance function*

$$g : [0,1]^{N_S} \times \cdots \times [0,1]^{N_S} \times \Omega \to \mathbb{R}_{\geq 0} \tag{3.6}$$

is introduced. The function g is used to determine the "distance" between the data $\pi(t,j)$ and the model approximation of the data, determined with the parameters $\theta(\bar{u}(t,j))$. Hence any metric $d(\cdot,\cdot)$ induces an appropriate fitness function

$$\begin{aligned} &g(\pi(t+1,j),\ldots,\pi(t-N_M,j),\theta(\bar{u}(t,j))) \\ &\quad = \Big(d(\pi(t+1,j),\mathbf{E}[f(\pi(t,j),\ldots,\pi(t-N_M,j),\theta(\bar{u}(t,j)))])\Big)^2. \end{aligned} \tag{3.7}$$

For instance, the model distance function example

$$g(\pi(t,j),\theta(t,j)) := \|\pi(t,j) - \theta(t,j)\|_2^2 \tag{3.8}$$

is derived from the Euclidean metric and based on the example model given in (3.2)[1]. The general inverse problem is phrased as follows:

$$\begin{aligned} &\mathbf{L}(\theta(\bar{u}(t,j))) \\ &\quad = \sum_{t=1}^{N_T}\sum_{j=1}^{N_J} g(\pi(t+1,j),\ldots,\pi(t-N_M,j),\theta(\bar{u}(t,j))) \to \min_{\theta(\bar{u}(t,j))}. \end{aligned} \tag{3.9}$$

For the logistic direct mathematical model function in (3.3), the corresponding inverse problem is defined as

$$\mathbf{L}(B(t,j)) = \sum_{t=1}^{N_T}\sum_{j=1}^{N_J} \Big\|\pi(t,j) - \theta^{\text{logit}}(B(t,j),u(t,j))\Big\|_2^2 \to \min_{B(t,j)}. \tag{3.10}$$

The reader is referred to [31, 84] for a discussion on various other direct model function examples. The focus in this thesis is on a particular Markov model function that allows to incorporate all implicit external factors driving the considered system. The joint impact of these unresolved factors is reflected in an explicit dependency on time and location. This Markov model is presented and discussed in the following subsection and will be used to identify arctic sea ice dynamics in Chapter 5.

1 Note that a potentially costly computation of the square root function is avoided via using the square of the Euclidean distance.

3.1.1 *Markov model*

Considering the time-wise dynamics of a process $\sigma(t,j,l)$ with a discrete state space, a typical approach is to assume that the probability of the current state depends only on the time-wise previous state. This property is called the *Markov property* and a process $\sigma(t,j,l)$ is referred to as (time-wise) Markovian if it fulfills the following condition:

$$\begin{aligned}\mathbb{P}[\sigma(t,j,l) = s_i | \sigma(t-1,j,l) = s_{h_{t-1}}, \dots, \sigma(1,j,l) = s_{h_1}] \\ = \mathbb{P}[\sigma(t,j,l) = s_i | \sigma(t-1,j,l) = s_{h_{t-1}}] \qquad \forall j,l.\end{aligned} \tag{3.11}$$

Markov chains in general [15] have been deeply studied and the corresponding theory is employed in various applicational areas, e.g., seasonal forecast of antarctic sea ice in the context of climatology [21] and for the analysis of political opinion polls in Germany in the field of computational sociology [54]. Suppose, a considered dynamical process $\sigma(t,j,l)$ with discrete state space driven by external factors $\bar{u}(t,j)$ is given by means of ensemble data $\pi(t,j)$ and has the Markov property. Then the process can fully be described with a transition matrix $P(\bar{u}(t,j))$ and the corresponding *stochastic master equation*

$$\pi(t+1,j)^T = \pi(t,j)^T P(\bar{u}(t,j)). \tag{3.12}$$

The entries of the matrix $P(\bar{u}(t,j))$ contain the transition probabilities:

$$\{P(\bar{u}(t,j))\}_{nm} = \mathbb{P}[\sigma(t,j,l) = s_m | \sigma(t-1,j,l) = s_n]. \tag{3.13}$$

As already discussed in Subsection 2.1.1, it is usually not possible to have access to all external factors $\bar{u}(t,j)$. Thus, the aim is to address the problem by designing a model specifically taking unknown external factors, i.e., $u^{\text{unres}}(t,j)$, into account. Essentially, the unresolved quantities are represented in form of a joint impact via an explicit dependency on time and space. Along the lines of [31], this is achieved by assuming a certain structure for the transition matrix. The details of this approach are discussed in the following proposition.

Proposition 3.1.1. *Let $P : \mathbb{R}^{N_F} \to \mathbb{R}^{N_S \times N_S}$ be a twice differentiable function with bounded second derivatives, then $P(\bar{u}(t,j))$ can be expressed by the following decomposition*

$$P(\bar{u}(t,j)) = P_0(t,j) + \sum_{e=1}^{N_E} P_e(t,j) u_e(t,j) + \varepsilon(t,j), \tag{3.14}$$

where the expected value of the noise process $\varepsilon(t,j)$ is equal to zero, i.e, $\mathbf{E}[\varepsilon(t,j)] = 0$, and $P_e(t,j) \in \mathbb{R}^{N_S \times N_S}\ \forall t,j$ denote time- and location-dependent matrices.

Proof. As P is a twice differentiable function with bounded second derivatives, it can be approximated with a Taylor-expansion. More specifically, a Taylor-expansion around the means

$$\mu(t,j) = [\mathbf{E}(\bar{u}_1(t,j)), \ldots, \mathbf{E}(\bar{u}_{N_E+N_I}(t,j))] \in \mathbb{R}^{(N_E+N_I)\times 1} \tag{3.15}$$

is considered, i.e,

$$\begin{aligned} P(\bar{u}(t,j)) = P(\mu(t,j)) &+ \sum_{e=1}^{N_E} \frac{\partial P(\mu(t,j))}{\partial \bar{u}_e(t,j)} (\bar{u}_e(t,j) - \mu_e(t,j)) \\ &+ \sum_{|\alpha|=2} R_\alpha(\bar{u}(t,j))(\bar{u}(t,j) - \mu(t,j))^\alpha, \end{aligned} \tag{3.16}$$

where the remainder term is defined as

$$R_\alpha(\bar{u}(t,j)) = \frac{2}{\alpha!} \int_0^1 (1-x) D^\alpha P(\mu(t,j) + x(\bar{u}(t,j) - \mu(t,j)))\, dx \tag{3.17}$$

with α being a multi-index. It is important to mention that $R_\alpha(\bar{u}(t,j))$ is bounded as the second derivatives of $P(\bar{u}(t,j))$ are assumed to be bounded. Further, note that the vector of external factors $\bar{u}(t,j)$ is without loss of generality assumed to be ordered starting with the resolved quantities (see (2.9)). The structure given in (3.14) is achieved by defining the matrices as follows:

$$P_e(t,j) := \frac{\partial P(\mu(t,j))}{\partial \bar{u}_e(t,j)} \quad \text{for } e \in \{1, \ldots, N_E\}, \tag{3.18}$$

$$\begin{aligned} P_0(t,j) := P(\mu(t,j)) &- \sum_{e=1}^{N_E} \frac{\partial P(\mu(t,j))}{\partial \bar{u}_e(t,j)} \mu_e(t,j) \\ &+ \mathbf{E}\left[\sum_{e=N_E+1}^{N_E+N_I} \frac{\partial P(\mu(t,j))}{\partial \bar{u}_e(t,j)} (\bar{u}_e(t,j) - \mu_e(t,j)) \right. \\ &\left. + \sum_{|\alpha|=2} R_\alpha(\bar{u}(t,j))(\bar{u}(t,j) - \mu(t,j))^\alpha \right]. \end{aligned} \tag{3.19}$$

The error process is set to be

$$\begin{aligned}\varepsilon(t,j) := &\sum_{e=N_E+1}^{N_E+N_I} \frac{\partial P(\mu(t,j))}{\partial \bar{u}_e(t,j)}(\bar{u}_e(t,j)-\mu_e(t,j)) \\ &+ \sum_{|\alpha|=2} R_\alpha(\bar{u}(t,j))(\bar{u}(t,j)-\mu(t,j))^\alpha \\ &- \mathbf{E}\left[\sum_{e=N_E+1}^{N_E+N_I} \frac{\partial \bar{P}(\mu(t,j))}{\partial \bar{u}_e(t,j)}(\bar{u}_e(t,j)-\mu_e(t,j))\right. \\ &\qquad \left. + \sum_{|\alpha|=2} R_\alpha(\bar{u}(t,j))(\bar{u}(t,j)-\mu(t,j))^\alpha\right]\end{aligned} \tag{3.20}$$

so that $\mathbf{E}[\varepsilon(t,j)] = 0$ immediately follows. Finally, resorting of the terms in (3.16) yields (3.14). □

Note that an additional assumption concerning the statistical independence of $(\bar{u}_e(t,j) - \mu_e(t,j))$ for all e, t, and j, implies that the different realizations of $\varepsilon(t,j)$ are independent of each other in j and t as well[2]. Due to $\mathbf{E}[\varepsilon(t,j)] = 0$, it is reasonable to assume that $\varepsilon(t,j)$ is small, yet its variance can take any value which might lead to arbitrary error values.

A variation of this proposition can be found in [31]. It states that $P(\bar{u}(t,j))$ can also be decomposed in a similar fashion considering the next order Taylor-expansion. The main structural difference is that the decomposition contains an additional and conceptually different (i.e., multiplicative) error term.

The following linear combination of the matrices $P_0(t,j)$ and $P_e(t,j)$ and the resolved external factors $u_e(t,j) \in \mathbb{R}$ with $e \in \{1,\dots,N_E\}$ is denoted

$$P(t,j,u(t,j)) := P_0(t,j) + \sum_{e=1}^{N_E} P_e(t,j)u_e(t,j). \tag{3.21}$$

Concluding, assuming the transition matrix $P(\bar{u}(t,j))$ is a twice differentiable function with bounded second derivatives, it is possible to approximate the observed dynamics of the process $\sigma(t,j,l)$ with the model parameter $P(t,j,u(t,j))$, dependent on explicit external factors $u(t,j)$, time t, and location j. The corresponding model function f with output $\pi(t+1,j)$ is induced by the stochastic master equation (see (3.12)), i.e.,

$$f^{\text{Markov}}(\pi(t,j), P(t,j,u(t,j))) = \pi(t,j)^T(P(t,j,u(t,j)) + \varepsilon(t,j)), \tag{3.22}$$

2 Note that this does not necessarily imply that $\varepsilon(t,j)$ is also identically distributed for all j and t.

where $\varepsilon(t,j)$ is the error term defined in Proposition 3.1.1 with $\mathbf{E}[\varepsilon(t,j)] = 0$. The model distance function is again derived from the Euclidean metric:

$$\begin{aligned} &g(\pi(t+1,j), \pi(t,j), P(t,j,u(t,j))) \\ &\quad = \left\| \pi(t+1,j)^T - \pi(t,j)^T P(t,j,u(t,j)) \right\|_2^2. \end{aligned} \tag{3.23}$$

Consequently, the associated inverse problem is defined as

$$\begin{aligned} &\mathbf{L}(P(t,j,u(t,j))) \\ &\quad = \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} \left\| \pi(t+1,j)^T - \pi(t,j)^T P(t,j,u(t,j)) \right\|_2^2 \rightarrow \min_{P(t,j,u(t,j))}. \end{aligned} \tag{3.24}$$

As the total number of unknown parameters is too large in comparison with the total number of data points (i.e., $N_T N_J$), the inverse problems needs to be regularized. The details are discussed in the following section.

3.2 INTERPOLATION

The current formulation of the underlying inverse problem given in (3.9) is ill-posed. More specifically, due to the explicit dependency of the model parameter $\theta(t,j,\bar{u}(t,j))$ on time and space, model inference of the considered inverse problem may result in model overfitting. For instance, this effect is visible considering the simple model function example given in (3.8). In that case the global optimum $\theta^*(t,j)$ is equivalent to the current state probability distribution, i.e.,

$$\theta^{\text{kmeans}*}(t,j) = \pi(t,j). \tag{3.25}$$

Note that existing global optima for a considered optimization problem are denoted with a superscript asterisk from here on. The trivial solution given in (3.25) to the general inverse problem (3.9) for the model function given in (3.2) does not provide any new information. Put differently, the great number of unknown model parameters relative to the number of observational data points (i.e., $N_T N_J$) causes the problem to be ill-posed according to the definition of Hadamard [45] and the results to be meaningless. In order to

address this issue, local stationarity and homogeneity are assumed. Thus, the model distance functional is reformulated, i.e.,

$$\begin{aligned} &g(\pi(t+1,j),\dots,\pi(t-N_M,j),\theta(\bar{u}(t,j))) \\ &\quad=\sum_{k=1}^{N_K}\gamma_k(t,j)g(\pi(t+1,j),\dots,\pi(t-N_M,j),\theta_k(u(t,j))), \end{aligned} \tag{3.26}$$

where

$$\Theta(u(t,j))=[\theta_1(u(t,j)),\dots,\theta_{N_K}(u(t,j))] \tag{3.27}$$

are local regimes independent of time and space and

$$\Gamma(t,j)=[\gamma_1(t,j),\dots,\gamma_{N_K}(t,j)]\in[0,1]^{1\times N_K} \tag{3.28}$$

is a jump process assigning affiliations to the N_K local model parameters $\theta_k(u(t,j))$ for all time steps t and locations j. Furthermore, the affiliation process $\Gamma(t,j)$ is subject to two convexity constraints:

$$\sum_{k=1}^{N_K}\gamma_k(t,j)=1 \quad \text{for } j\in\{1,\dots,N_J\}, t\in\{1,\dots,N_T\}, \tag{3.29}$$

$$\gamma_k(t,j)\geq 0 \quad \text{for } j\in\{1,\dots,N_J\}, t\in\{1,\dots,N_T\}, k\in\{1,\dots,N_K\} \tag{3.30}$$

and it can also be interpreted as a path switching between different clusters. Hence the adaptive version

$$\mathbf{L}(\Gamma,\Theta)=\sum_{j=1}^{N_J}\mathbf{L}_j(\Gamma(:,j),\Theta(u(t,j)))\rightarrow\min_{\Gamma(t,j),\Theta(u(t,j))} \tag{3.31}$$

with

$$\begin{aligned} &\mathbf{L}_j(\Gamma(:,j),\Theta) \\ &\quad=\sum_{t=1}^{N_T}\sum_{k=1}^{N_K}\gamma_k(t,j)g(\pi(t+1,j),\dots,\pi(t-N_M,j),\theta_k(u(t,j))) \end{aligned} \tag{3.32}$$

of the functional $\mathbf{L}(\Theta(t,j,u(t,j)))$ is denoted the *average clustering functional*. Consequently, the exemplary inverse problem for an a priori assumed logistic model is changed to

$$\begin{aligned} \mathbf{L}(\Gamma,B_1,\dots,B_{N_K})&=\sum_{t=1}^{N_T}\sum_{j=1}^{N_J}\gamma_k(t,j)\left\|\pi(t,j)-\theta^{\text{logit}}(B_k,u(t,j))\right\|_2^2 \\ &\rightarrow\min_{\Gamma(t,j),B_1,\dots,B_{N_K}} \end{aligned} \tag{3.33}$$

where the local stationary and homogenous model parameters are given by the vector $B_k = \left[B_k^1, \ldots, B_k^{N_S}\right]$ for all $k \in \{1, \ldots, N_K\}$. The average clustering functional corresponding to the Markov model function, introduced in Subsection 3.1.1 (see (3.24)), can be expressed via

$$\begin{aligned} &\mathbf{L}(\Gamma(t,j), P(u(t,j))) \\ &= \sum_{j=1}^{N_J} \sum_{t=1}^{N_T} \sum_{k=1}^{N_K} \gamma_k(t,j) \left\| \pi(t+1,j)^\top - \pi(t,j)^\top P^k(u(t,j)) \right\|_2^2 \to \min_{\Gamma(t,j), P(u(t,j))}. \end{aligned} \tag{3.34}$$

The model matrices $P^k(u(t,j))$ characterizing the different regimes are defined as

$$P^k(u(t,j)) = P_0^k + \sum_{e=1}^{N_E} P_e^k u_e(t,j) \quad \forall\, k \in \{1, \ldots, N_K\}. \tag{3.35}$$

Thus, the matrix structure proposed in Proposition 3.1.1 (see (3.21)) is maintained for each of the local model matrices. The corresponding vector of stationary, homogenous model matrices is denoted

$$P(u(t,j)) := \left[P^1(u(t,j)), \ldots, P^{N_K}(u(t,j))\right] \in \mathbb{R}^{N_S \times N_S N_K}. \tag{3.36}$$

In order to ensure the stochasticity of the model transition matrix, the minimization problem is subject to the following constraints:

$$P_0^k \mathbf{1} = \mathbf{1} \quad \forall\, k \in \{1, \ldots, N_K\}, \tag{3.37}$$

$$P_e^k \mathbf{1} = \mathbf{0} \quad \forall\, e \in \{1, \ldots, N_E\},\ k \in \{1, \ldots, N_K\}, \tag{3.38}$$

where $\mathbf{1} \in \mathbb{R}^{N_S \times 1}$ is the column vector with all entries equal to one and analogously $\mathbf{0} \in \mathbb{R}^{N_S \times 1}$ refers to the corresponding vector with all entries equal to zero. Moreover, the model matrices $P^k(u(t,j))$ are required to be (element-wise) non-negative for all t, j and k, i.e.,

$$\left\{P^k(u(t,j))\right\}_{n,m} \geq 0 \quad \forall\, t, j, k \text{ and } n, m \in \{1, \ldots, N_S\}. \tag{3.39}$$

However, computational difficulties arise in the context of implementing this additional condition. More specifically, due to the dependency of $P^k(u(t,j))$ on the vector of explicit external factors $u(t,j)$, the numerical costs to ensure constraint (3.39) are immense. Consequently, the constraint is computationally unfeasible. Yet, it is possible to ensure

$$\left\{P_0^k\right\}_{n,m} \geq 0 \quad \forall\, k, n, m. \tag{3.40}$$

due to the fact that P_0^k is independent of $u(t,j)$. Further, as demonstrated in [84], the computational complexity can be reduced, assuming that the convex hull of the space $\mathcal{U}$ containing the vector of explicit external factors $u(t,j)$ is a N_E-dimensional hypercube. Then the non-negativity of the model matrices (see (3.39)) is fulfilled if

$$\sum_{e=1}^{N_E} \left\{ P_e^k \right\}_{n,m} \begin{bmatrix} \sup\limits_{t,j} u_e(t,j) \\ \inf\limits_{t,j} u_e(t,j) \end{bmatrix} \geq 0 \quad \forall\, k,n,m. \tag{3.41}$$

Summarizing, the priorly computationally complex problem can be reduced to 2^{N_E} inequality constraints. Details concerning the implementation of the considered numerical problem can be found in Appendix A.1.

In order to gain a better understanding of the underlying process $\sigma(t,j,l)$, a special case of the considered Markov model is considered in the following subsection.

3.2.1 *Special case: Memory-less process*

Suppose, the real life dynamical process $\sigma(t,j,l)$ under consideration does in fact not depend on the previous state probabilities $\pi(t-1,j)$, i.e, is a memory-less process. In other words, the regarded system can be approximated with a model function similar to the one defined in (3.2). Essentially, such an independent process is a special case of the considered Markov model. More precisely, it is possible to consider the direct model function given in (3.22) with the transition matrix structure given in (3.35) subject to the additional constraint that the columns entries of the matrices P_e^k have to be equal for all e and k, i.e.,

$$\left\{ P_e^k \right\}_{1,m} = \left\{ P_e^k \right\}_{2,m} = \cdots = \left\{ P_e^k \right\}_{N_S,m}, \quad m \in \{1,\ldots,N_S\},\ \forall\, e,k \tag{3.42}$$

Then, due to the fact that the entries of the state probabilities sum up to one, future state probabilities $\pi(t+1,j)$ can be approximated independently of the current distribution $\pi(t,j)$. By distinguishing between the standard Markov model and this independent special case, it is possible to gain a deeper insight into the dynamics underlying the considered data. Therefore, the parametrization of the arctic sea ice dynamics in Chapter 5 is done for a general Markov model as well as for the independent special case. A model selection criterion (see Section 3.6) will be used to determine the model better suited to describe the considered system.

3.3 SPATIAL AND TEMPORAL PERSISTENCE

Unfortunately, the average clustering functional introduced in (3.31) is still ill-posed in the sense of Hadamard [45]. More specifically, for a fixed vector $\Theta(u(t,j)) = [\theta_1(u(t,j)), \dots, \theta_{N_K}(u(t,j))]$, the optimal regime assigning $\Gamma^*(t,j)$ with respect to the model distance function g is

$$\gamma_k^*(N_T+1,j) \tag{3.43}$$
$$= \begin{cases} 1 & \text{if } k = \underset{h}{\operatorname{argmin}}\, g(\pi(t+1,j), \dots, \pi(t-N_M,j), \theta_h(u(t,j))), \\ 0 & \text{otherwise.} \end{cases}$$

The affiliation $\gamma_k^*(N_T+1,j)$ can exhibit discontinuous behavior, and in those cases the persistency assumption of the previous section is violated. Thus, further constraints need to be imposed on $\Gamma(t,j)$ in order to prevent the process from rapidly switching between regimes. Two different regularization approaches have been proposed [52, 53, 54] to enforce persistency on $\Gamma(t,j)$. In the following, each ansatz is briefly summarized, and the reader is referred to [84] for a detailed discussion.

3.3.1 *Tikhonov regularization*

To improve the posedness of the minimization problem (formulated in (3.31)), the function space of Γ is restricted and a Tikhonov regularization [110] is deployed. Tikhonov regularizations are frequently applied in the context of image processing [123] and ill-posed interpolation problems [117] and have recently been proposed for clustering problems of the form given in (3.31) [52].

The affiliation processes $\gamma_k(:,j)$ are assumed to be weakly differentiable, i.e.,

$$\gamma_k(:,j) \in W^{1,2}([1,N_T]) \quad \forall j,k, \tag{3.44}$$

where $W^{1,2}([1,N_T])$ is the Sobolev space containing all real functions of $L^2([1,N_T])$ who's first weak derivative also belongs to $L^2([1,N_T])$. The additional information concerning the path space of Γ allows to phrase a modified average clustering functional, namely

$$\mathbf{L}^{\mathfrak{r}}(\Gamma,\Theta) = \sum_{j=1}^{N_J} \mathbf{L}_j^{\mathfrak{r}}(\Gamma(:,j),\Theta) \rightarrow \min_{\gamma_k(:,j)\in W^{1,2}([1,N_T])\ \forall\ k,j,\Theta} \tag{3.45}$$

with

$$\mathbf{L}_j^{\mathfrak{r}}(\Gamma(:,j),\Theta) = \mathbf{L}_j(\Gamma(:,j),\Theta) + \mathfrak{r}^2 \sum_{k=1}^{N_K} \left\| \frac{\partial \gamma_k}{\partial t} \right\|^2_{L^2([1,N_T])}, \tag{3.46}$$

where the norm $\left\| \frac{\partial \gamma_k}{\partial t} \right\|^2_{L^2([1,N_T])}$, measuring the smoothness of the path $\gamma_k(\cdot, j)$, is defined as

$$\left\| \frac{\partial \gamma_k}{\partial t} \right\|^2_{L^2([1,N_T])} = \int_1^{N_T} \left(\frac{\partial \gamma_k}{\partial t} \right)^2 dt \quad \forall j,k. \tag{3.47}$$

This modified version of the average clustering functional is referred to as *Tikhonov-regularized average clustering functional*. An optimal process $\Gamma^*(t,j)$ minimizing the Tikhonov-regularized version (3.45) of the average clustering functional can be smoothed by increasing the value of the regularization factor $\mathfrak{r}$. This effect has been studied for several synthetic as well as real data sets of different applicational fields in [30, 52].

The advantage of the introduced Tikhonov ansatz is that no additional constraints are imposed on the affiliation process $\Gamma(t,j)$ making it possible to apply Markov chain Monte Carlo techniques (MCMC) to find an optimal $\Gamma^*(t,j)$ for $\mathbf{L}^{\mathfrak{r}}(\Theta,\Gamma)$ for a fixed $\Theta(u(t,j))$. An MCMC-minimization of this clustering problem is outlined in Subsection 3.5.1.

Although the process $\Gamma^*(t,j)$ can be smoothed to a certain degree by carefully tuning the regularization factor $\mathfrak{r}$, no direct control over the number of transitions between the regimes can be gained with the Tikhonov regularization. An alternative approach addressing this issue via adding a persistency constraint has been proposed in [53, 54] and is outlined in the next subsection.

3.3.2 *BV-regularization*

Following [53, 54], the process $\gamma_i(:,j)$ is assumed to be a function with bounded variation. Note that this includes all the functions contained in $W^{1,2}([1,N_T])$ as well as functions with discontinuities, e.g., jumps. Consequently, it is possible to gain direct control over the persistency of $\gamma_i(\cdot,j)$ by adding the constraint

$$|\gamma_k(:,j)|_{BV(1,N_T)} = \sum_{t=1}^{N_T-1} |\gamma_k(t+1,j) - \gamma_k(t,j)| \leq N_C \tag{3.48}$$

for all locations $j \in \{1,\ldots,N_J\}$. Concluding, Condition (3.48) ensures that the maximal number of transitions has upper bound N_C. Due to the fact that

the constraints for $\Gamma(t, j)$ are independent for every location, it is possible to compute each $\Gamma(:, j)$ separately for fixed $\Theta(u(t, j))$. Thus, finding an optimal $\Gamma^*(t, j)$ is a *linear minimization problem with linear constraints*. The details of the numerical approach to compute a minimizing affiliation process are discussed in [84]. Nevertheless, it is important to stress that due to the additional constraint, a Markov chain Monte Carlo optimization ansatz is not an option for this regularization.

An analog approach to limit the total number of transitions between the different regimes along all locations j for a fixed t might be worthwhile to describe systems under consideration but presents an immense computational challenge. These big computation costs are a result of a then necessary global coupling (in j) for different optimization problems $\mathbf{L}_j$. Thus, an additional constraint restricting the locations remains an aspect of further research.

3.4 SPATIAL RELATIONS

As the data has a time as well as a spatial component, it is important to consider possible correlations or interactions between certain locations. A lot of effort has been put into the development of parametrization tools that allow to simultaneously model existing time-wise and spatial relations.

In the context of classical time series analysis, many standard methods that are able to accurately approximate the system underlying time-dependent data are not equipped to capture existing spatial correlations. Vice versa, this holds true for many of the techniques primarily developed for a spatial component. Consequently, purely time-dependent data is often approached with regression analysis or other members of the family of generalized linear models [80] where dynamics underlying observations with primarily spatial components are described with models such as variograms [26] and Markov random fields [65].

Summarizing, the research has been focusing on either time- or location-dependent data sets and, typically, phenomena associated with spatial expansion (respectively time-evolution) were discarded in order to concentrate on the main problem at hand.

As computational devices improved and higher dimensional data sets can be handled, the focus has shifted and both dimensions are considered simultaneously. Thus, the scientific community is presented with the new challenge of extending existing methodologies or designing entirely new frameworks fitting both dimensions.

The introduced Markov model was first proposed for purely time-dependent data [53] and later modified to fit time series with a dependency on location as well. As demonstrated in [31], spatial relations can be included in the current non-stationary, non-homogenous Markov model by adding an explicit external factor

$$u_{N_E+1}(t,j) := \underset{\mathrm{r}\in\ \mathrm{neigh}(j)}{\mathbf{average}}(\pi(t-1,r)) \tag{3.49}$$

to the vector of measurable exterior forces $u(t,j)$, where $\mathrm{neigh}(j)$ denotes the set of all direct neighboring locations of a cell j. Note that the exact definition of $\mathrm{neigh}(j)$ highly depends on the lattice, which should be chosen in accordance with the application.

Summarizing, the mean of the time-wise previous state probabilities of the neighboring cells is considered in order to contemplate potentially existing spatial correlations. The importance of these interactions and influences of neighboring cells is revisited in Subsection 5.3.3, where the statistical impact of adjacent location states is evaluated by means of a considered arctic sea ice application. Note that in order to get an even deeper understanding of the interactions between neighboring cells one could include further information, such as the discrete gradients of the previous state probabilities in the vector of explicit external factors.

3.5 NUMERICAL APPROACH AND COMPUTATIONAL COMPLEXITY

Unfortunately, the inverse problem posed in (3.31) is not convex. Consequently, it can not be anticipated to obtain a global minimum with commonly deployed techniques such as gradient descent or Newton methods. Nevertheless, following [52], the problem of approximating global minimizers $\Gamma^*(t,j)$ and $\Theta^*(u(t,j))$ can be addressed combining a *subspace algorithm* with a simulated annealing ansatz [66].

Instead of simultaneously minimizing $\mathbf{L}(\Gamma,\Theta)$ for both unknown parameters Γ and Θ, the conceptual idea is to exploit the structure of the average clustering functional and divide the inverse problem into two minimizations over just one parameter. In particular, this means to optimize $\mathbf{L}(\Gamma,\Theta)$ with respect to Γ for a fixed Θ and, vice versa, with respect to Θ for a fixed Γ which can be done with standard optimization techniques (e.g., simplex method [28, 29, 122])).

Subsequent iterations over the sub optimization problems allow to determine local minima for the model parameters but there is no guarantee to compute a global optimum. Thus, the subspace algorithm itself is repeated

in order to find the global minima. For a large number of repetitions this standard simulated annealing approach [66, 71] is deployed to avoid calculating local minima that are not global optimal solutions. Although this form of the simulated annealing concept is well established, there is still no guarantee that a global minimum is found. The details of the procedure are explained in Algorithm 1 by means of the optimization problem (3.34) with constraints.

Before executing the algorithm, it is necessary to choose the values of the free variables, such as the memory $N_M \in \{0, 1\}$ of the process $\sigma(t, j, l)$, the number of local stationary and homogenous regimes, and N_C the upper bound for the transitions of the regime assigning process Γ. Note that, for $N_M = 1$, a Markov process with memory and, for $N_M = 0$, an independent process is assumed. In detail that means that constraint (3.42) is switched off or on for the computations of a minimal $P(u(t, j))$ for a fixed affiliation process $\Gamma(t, j)$ (see Step 2 of Algorithm 1).

Moreover, it is important to select the vector $u(t, j)$ of explicit external factors and the data $\pi(t, j)$ associated with the considered dynamical process $\sigma(t, j, l)$. Further, computational settings such as the number of annealing steps $N_{\text{anneal}}^{\text{FEM}}$ and the numerical tolerance $N_{\tau\text{ol}}^{\text{FEM}}$ for the subspace iterations have to be fixed.

Another important choice concerns the considered data product. Especially the size of the vector of explicit external factors and the variety of its entries needs to be considered. Details on the specific aspects of this particular choice for the considered application of analyzing the dynamics of the arctic sea ice concentration are discussed in Chapter 5.

The output of the algorithm is a model

$$\mathcal{M}^f(N_K, N_C, N_M, u(t, j))) \tag{3.50}$$

consisting of global optimizers $\Gamma^*(t, j)$ and $P^*(u(t, j))$ which depend on N_K, N_C, and N_M. The terminology $P^{[s]}(u(t, j))$ and $\Gamma^{[s]}(t, j)$ denotes the current approximations of the optimal $P^*(u(t, j))$ and $\Gamma^*(t, j)$.

The optimization of $\mathbf{L}_j$ with respect to $\Gamma(:, j)$ (see Lines 6-8) of the BV-regularized averaged clustering functional can be approached with standard methods of linear minimization with linear equality and inequality constraints (e.g., simplex method [28, 29, 122]). As the affiliation process Γ is not subject to any spatial persistency constraints, the problem of optimizing $\mathbf{L}$ with respect to Γ for fixed $P(u(t, j))$ is equivalent to separate computations of the individual sub-optimization problems $\mathbf{L}_j$ (see (3.32)) with respect to

Algorithm 1: Subspace algorithm with annealing steps

input :

- Data: $\pi(t,j)$, $u(t,j)$
- Model variables: N_C, N_K, N_M
- Numerical settings: $N_{\tau\text{ol}}^{\text{FEM}}$, $N_{\text{anneal}}^{\text{FEM}}$
- Optional: $N_{\text{basis}}^{\text{FEM}}$

output:

- $\Gamma^*(t,j)$ and $P^*(u(t,j))$

$\mathbf{L}_{min} = 1000000$
for $r = 1 : N_{anneal}^{FEM}$ **do**
  Generate random initial $\Gamma^{[0]}(t,j)$ *and compute* $P^{[0]}(u(t,j))$
  $s = 1$
  while $|\mathbf{L}(\Gamma^{[s]}(t,j), P^{[s]}(u(t,j))) - \mathbf{L}(\Gamma^{[s-1]}(t,j), P^{[s-1]}(u(t,j)))| \geq N_{\tau ol}^{FEM}$ **do**
    Step 1:
    for $j = 1 : N_J$ **do**
      Determine $\Gamma^{[s+1]}(:,j) = \arg\min \mathbf{L}_j(\Gamma(:,j), P^{[s]}(u(t,j)))$ *subject to constraints (3.29),(3.30)*
    Step 2:
    Compute $P^{[s+1]}(u(t,j)) = \arg\min \mathbf{L}(\Gamma^{[s+1]}, P(u(t,j)))$ *subject to constraints (3.37), (3.38), (3.40), and (3.41) (for* $N_M = 0$ *condition (3.42) also needs to be fulfilled)*
    $s := s + 1$
  if $\mathbf{L}_{min} \geq \mathbf{L}(\Gamma^{[s]}(t,j), P^{[s]}(u(t,j)))$ **then**
    $\mathbf{L}_{min} = \mathbf{L}(\Gamma^{[s]}(t,j), P^{[s]}(u(t,j)))$
    $\Gamma^* = \Gamma^{[s]}(t,j)$
    $P^*(u(t,j)) = P^{[s]}(u(t,j))$

$\Gamma(:,j)$ for all locations $j \in \{1,\ldots,N_J\}$. Details concerning the implementation are discussed in [84].

In general, the run time of Step 1 scales with the dimension of $\Gamma(:,j)$ [84]. Thus, time-wise long time series can result in major computation times. In order to address this problem, a finite element approach was proposed in [52] to reduce the dimension of $\Gamma(:,j)$. The key idea is to exploit the persistency of the affiliation process. As the affiliations remain the same for a period of time, it is possible to find intervals $[i,i_n]$ so that $\Gamma(t,j)$ can be represented by $\Gamma(t_i,j)$ for all $t \in [t_i,t_{i_n}]$. Consequently, a discretized version

of Γ can be used in Step 1 and is enhanced to its original size N_T for the calculations of Step 2. Hence, the computational time of Step 1 only depends on the number of functions $N_{\text{basis}}^{\text{FEM}}$ (i.e, the size of the discretized Γ), which can be significantly smaller than the actual time dimension N_T for a very persistent path $\gamma_k(t, j)$.

In [30] the performance with respect to the total number of finite element functions $N_{\text{basis}}^{\text{FEM}}$ of a optimization via a simplex method was compared to the results of an MCMC optimization ansatz[3]. The Metropolis algorithm proved to be much faster, especially for the parametrization of dynamical systems that are accessible via data with a high dimensional time component and that have very persistent behavior. In other words, an MCMC optimization approach is a good alternative ansatz for dynamical processes that require a big number of finite element base functions to determine an appropriate corresponding Γ. However, due to the additional constraint relating to the BV-regularization, such an MCMC approach is only an option for a Tikhonov-regularized average clustering functional.

In general, the optimization of Θ for a fixed regime assigning process $\Gamma(t, j)$ depends on the previously chosen model function f. Here the focus is on a Markov model (see (3.22)). Thus, Step 2 of the subspace algorithm (see Lines 9-10) depicts the computation of $P^{[s+1]}(u(t, j))$ for the current $\Gamma^{[s+1]}(t, j)$.

The associated optimization problem is subject to constraints (3.37), (3.38), (3.40), and (3.41). It is possible to apply standard methods of quadratic optimization with linear equality and inequality constraints to find a minimal $P^*(u(t, j))$. Due to the Markov model choice, Algorithm 1 is also referred to as *non-stationary, non-homogenous Markov regression*.

In contrast to the separate individual computations that can be executed to calculate the optimal $\Gamma^*(:, j)$ for all locations j, the optimal local stationary and homogenous model parameters $P^k(u(t, j))$ with $k \in \{1, \ldots, N_K\}$ have to be computed simultaneously for all locations j and for all time steps t. Nevertheless, it is important to mention that the proposed framework can compete with standard approaches such as SVMs and ANNs regarding the computational complexity and the quality of the model approximation [30].

In order to fit an SVM model to the data, a quadratic minimization optimization problem has to be solved. For a gaussian RBF kernel function (see (1.10)) this can lead to a worst case complexity of $\mathcal{O}(N_T^2 N_E)$ for each location [12]. Yet, the mean computation time is usually much lower. In particular, tuning the regularization parameter $N_{\text{boxconstraint}}^{\text{SVM}}$ or increasing

3 In detail, the run time of Step 1 of Algorithm 1 was compared.

the size of the training data result in much faster convergence to a globally optimal solution [101].

In pursuance of calculating an appropriate feedforward network with a non-linear transfer function on the basis of data, a sequence of quadratic optimization problems has to be solved. Note that contrary to the unique robust solution, provided in the context of SVMs, the network is fitted via a non-convex gradient-based optimization method. Consequently, annealing steps, which increase the computation time, are necessary. Even for very efficient methods such as the Levenberg-Marquardt backpropagation algorithm [46] the run time scales badly with the total number of involved parameters. Here the total number of unknown variables is the sum of the parameters required for each neuron (considering all layers), where the number of parameters of a neuron is the sum of the total number of weights (entries of vector $\mathcal{W}$) plus one (due to the bias $\mathfrak{b}$).

The numerical details of Step 2 for a Markov model as proposed in (3.22) are discussed in Appendix A. A C++ code of optimization problem (3.34) for a fixed regime assigning process Γ has been implemented and can be found on `http://www.dewiljes.de/dewiljes/Jana.html`. Further, an open source quadratic programming solver[4] is employed for the computations in Chapter 5.

Note that only the BV-regularized variant was considered in Algorithm 1. Yet, the subspace algorithm for the minimization of a Tikhonov-regularized average clustering functional is conceptually analog. As demonstrated in [52], a Tikhonov-regularized average clustering functional for fixed Θ can be minimized with respect to Γ with standard quadratic optimization tools [42]. An alternative option is to employ MCMC-optimization via sampling from an appropriately chosen Boltzmann distribution [22]. One advantage of such a stochastic optimization approach is that the run time can be improved considerably. This ansatz is outlined in Subsection 3.5.1 and a detailed discussion can be found in Appendix B.

3.5.1 *MCMC approach*

The main computational drawback of the regularized model distance functional $\mathbf{L}^{\iota}(\Gamma, \Theta)$ as well as $\mathbf{L}(\Gamma, \Theta)$ is that they are non-convex. Consequently, it is not possible to ensure that any results determined with standard optimization techniques, such as simplex methods [28, 29], are global optima.

4 The open source package can be downloaded on `http://www.diegm.uniud.it/digaspero/index.php?page=software`, the theory corresponding algorithm is discussed in [42].

In the subspace algorithm, described in Algorithm 1, simulated annealing steps in form of subsequent repetitions are executed to approach a global rather than a local minimum.

As demonstrated in [30, 37], this locality problem can also be directly addressed for a Tikhonov-regularized average clustering functional by employing a stochastic optimization technique for each of the individual suboptimization problems $\mathbf{L}_j^{\mathfrak{r}}(\Gamma(:,j))$. More specifically, the approach is based on sampling from an appropriately chosen corresponding Boltzmann distribution (also known as Gibbs measure), i.e.,

$$\mathcal{F}_{\mathbf{L}_j^{\mathfrak{r}},\beta}(\Gamma(:,j)) = \frac{1}{Z(\mathbf{L}_j^{\mathfrak{r}})} \exp(-\beta \mathbf{L}_j^{\mathfrak{r}}(\Gamma(:,j),\Theta)) \tag{3.51}$$

for fixed model parameter Θ and fixed location j. The corresponding normalizing constant is defined as

$$Z(\mathbf{L}_j^{\mathfrak{r}}) = \int_{\Gamma} \exp(-\beta \mathbf{L}^{\mathfrak{r}}(\Gamma,\Theta)). \tag{3.52}$$

As the Boltzmann distribution has its origin in statistical physics, where it is used to describe certain phenomena in the field of thermodynamics, the parameter β is referred to as the *inverse temperature* and the regarded functional $\mathbf{L}^{\mathfrak{r}}$ is referred to as the *energy* of the considered system. More specifically, the Gibbs measure can be employed to describe the probability of a particle's speed influenced by the external temperature.

In particular, Boltzmann distributed samples Γ have the property to minimize the assigned energy functional $\mathbf{L}^{\mathfrak{r}}(\Gamma)$ as β converges to ∞ [47]. Thus, a Boltzmann distributed sample Γ is a good approximation of the global optimal Γ^*, minimizing (3.31).

However, it is usually not an option to calculate a normalizing constant $Z(\mathbf{L}^{\mathfrak{r}})$, such as given in Equation (3.52), since the high dimension of the path space hampers numerical computations. Yet, this problem can be avoided by employing techniques from the family of MCMC methods. More precisely, the Metropolis algorithm can be used as it is not necessary to determine $Z(\mathbf{L}^{\mathfrak{r}})$ in order to sample from the given Boltzmann distribution (3.51).

In the next paragraph, a basic introduction of a random walk Metropolis algorithm is given for the inverse problem phrased in (3.45). Further, the deployment of this specific MCMC technique to approximate Γ^* is explained.

METROPOLIS ALGORITHM First introduced in [83], the Metropolis algorithm is still frequently applied in various applicational areas and new

developments are steadily appearing in the current scientific literature [90]. The main advantage of the algorithm is that it allows to sample from distributions where a direct calculation is not possible as the boundaries of todays computational limits are reached.

For instance, the normalizing constant in the Boltzmann distribution, defined in (3.51), can not be computed directly, yet it is possible to sample from the Boltzmann distribution via the Metropolis algorithm. More specifically, numerical determination of the normalizing constant is not necessary as the Boltzmann distribution of an argument is only considered as a quotient of itself in a different argument, i.e.,

$$\frac{\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma')}{\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma)}. \tag{3.53}$$

The output of the considered Metropolis algorithm is a Markov chain of random samples, denoted $\Gamma^{[r]}$ with $r \in \{1, \ldots, N_{\text{chain}}^{\text{RWM}}\}$, that has the Gibbs measure as its unique stationary distribution[5]. An element $\Gamma^{[r]}$ of the chain, representing the current approximation of the global optimum Γ^*, is determined via an acceptance-rejection procedure, i.e, a potential new candidate Γ', generated with a *proposal density*

$$q(\cdot,\cdot), \tag{3.54}$$

is accepted or rejected depending on the *acceptance rate*

$$\begin{aligned} &\mathfrak{a}(\Gamma^{[r-1]},\Gamma') \\ &\quad= \begin{cases} \min\left\{1, \frac{\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma')q(\Gamma',\Gamma^{[r-1]})}{\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma^{[r-1]})q(\Gamma^{[r-1]},\Gamma')}\right\} & \text{if } \mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma^{[r-1]})q(\Gamma^{[r-1]},\Gamma') > 0, \\ 1 & \text{otherwise,} \end{cases} \end{aligned} \tag{3.55}$$

where $\Gamma^{[r-1]}$ is the previous chain member. If the acceptance rate is greater than a realization of a random variable[6], the proposed sample Γ' is accepted and becomes the new chain element $\Gamma^{[r]}$.

Different choices and specifications concerning the choice of the proposal density $q(\cdot,\cdot)$ lead to different variations of the original Metropolis algorithm. The original and standard principle, referred to as Random Walk Metropolis

5 Note that the samples can also follow a different target distribution, but in this thesis the regarded target distribution is the Boltzmann distribution with a Tikhonov-regularized average clustering functional as its energy function (see (3.51)).

6 The considered random variable follows a uniform distribution and takes values in $[0,1]$.

(RWM), is to update the previous chain member $\Gamma^{[r-1]}$ by adding a random *noise*, i.e.,

$$\Gamma(:,j)' = \Gamma^{[r-1]}(:,j) + \eta(:,j) \quad \forall j, \tag{3.56}$$

where

$$\eta(:,j) \sim q(\Gamma, \Gamma') \tag{3.57}$$

[22]. This simple, yet effective, strategy is just one option and there are many different strategies involving various families of proposal densities [73] to generate a new proposal sample, e.g., the more general Metropolis-Hastings algorithm [48], where each new candidate is generated independently of any previous chain element, or the Metropolis Adjusted Langevin Algorithm (MALA) [92, 90], where a new proposal depends on gradient information of the given optimization problem. Each ansatz has its advantages but, depending on the problem, can also present some challenges, e.g., independent walk Metropolis performs best when the proposal density is similar to the target distribution. Thus, it is sensible to choose the sample update considering the regarded problem.

Here, an RWM ansatz (see (3.56)) is proposed to generate new samples. More specifically, a new candidate Γ' is generated by adding a noise η, which is sampled from a Gaussian proposal density

$$\begin{aligned} &q\left(\Gamma^{[r-1]}(:,j), \Gamma(:,j)'\right) \\ &= \frac{1}{(2\pi)^{\frac{N_T}{2}}} \exp\left(-\frac{1}{2}\left(\Gamma(:,j)' - \Gamma^{[r-1]}(:,j)\right)^\top \left(\Gamma(:,j)' - \Gamma^{[r-1]}(:,j)\right)\right) \quad (3.58) \\ &= \frac{1}{(2\pi)^{\frac{N_T}{2}}} \exp\left(-\frac{1}{2}\eta(:,j)^\top \eta(:,j)\right), \quad (3.59) \end{aligned}$$

to the previous element $\Gamma^{[r-1]}$. This choice is rather basic but an RWM update has the advantage that no additional a priori information on the considered Boltzmann distribution is required. Further, using this specific generation of a new sample Γ', it is possible to directly control the acceptance rate $\alpha(\Gamma^{[r-1]}, \Gamma')$ which allows to employ an adaptive simulated annealing scheme used to improve the approximation of Γ^*. Moreover, due to the fact that the samples $\Gamma^{[r]}$ are subject to constraints in the regarded optimization problem, it is prudent not to involve gradient information.

STOCHASTIC OPTIMIZATION TO DETERMINE Γ In the following, a pseudocode describing the Metropolis algorithm is deployed to find $\Gamma^*(:,j)$ minimizing $\mathbf{L}_j^{\mathrm{r}}(\Gamma(:,j), \Theta)$, given in (3.46), for a fixed Θ. In other words, the Γ

optimization (see Step 1 of Algorithm 1) for a Tikhonov-regularized averaged clustering functional is approached with stochastic minimization.

The input of the algorithm includes the total number of different regimes N_K, the regularization factor $\mathfrak{r}$, the required maximal length of the Markov chain $\mathfrak{r}$, $N_{\text{chain}}^{\text{RWM}}$, the inverse temperature parameter β, and a fixed approximation of model parameter Θ.

The aim is to obtain an estimate of a global minimizer $\Gamma^*(:, j)$. Unfortunately, there is no direct option to include constraints restricting a sample. Thus, due to the additional constraint (see (3.48)), a BV-regularization can not be considered. Yet, the model parameter $\Gamma(:, j)$ has to satisfy conditions (3.29) and (3.30) for the Tikhonov-regularized inverse problem given in (3.45) (see Line 4 in Algorithm 2).

Suppose, the regarded process $\sigma(t, j, l)$ can be described with two regimes, i.e., $N_K = 2$, then the sampling procedure can be simplified. In particular, it is only necessary to sample a path $\gamma_1(:, j)$ subject to constraint (3.30), i.e, $\gamma_1(t, j) \in [0, 1]$, instead of sampling Γ with the additional condition (3.29)[7]. Since constraint (3.29) has to be fulfilled, the computation of the affiliation process corresponding to the second regime is straightforward, i.e,

$$\gamma_2(t, j) = 1 - \gamma_1(t, j) \quad \forall\, t, j. \tag{3.60}$$

However, this special case can not be regarded as a general solution to the problem. Following [30], it is possible to approximate a Boltzmann distributed and, therefore, optimal sample $\Gamma^*(:, j)$ satisfying (3.29) and (3.30) by assuming that the model parameter Γ depends on processes $\psi_k(t, j)$ and has the following analytic expression:

$$\gamma_k(\psi(t, j)) = \frac{\exp(\psi_k(t, j))}{\sum_{h=1}^{N_K} \exp(\psi_h(t, j))} \quad k \in \{1, \dots, N_K\}. \tag{3.61}$$

Thus, instead of directly sampling affiliations $\gamma_k(t, j)$, the RWM algorithm is used to sample with respect to the process $\psi(t, j) = [\psi_1(t, j), \dots, \psi_{N_K}(t, j)]$. Note that the acceptance rate $\mathfrak{a}(\Gamma^{[r-1]}(:, j), \Gamma'(:, j))$ is then computed for $\Gamma(\psi(:, j))$, defined in (3.61). Concluding, the new assumption allows to generate an approximation $\Gamma^*(:, j)$ that fulfills the required constraints for an arbitrary number of local regimes (i.e., $N_K \geq 3$).

7 To fulfill convexity constraint (3.30), which states that the entries of each of the new proposals Γ' (see (3.56)) have to be non-negative and to ensure that the entries are not greater than one the entries of the proposed sample Γ' are adjusted to suit the boundaries by setting the entries greater than one to one and the negative values to zero.

Algorithm 2: Metropolis algorithm

input :

- Data: $\pi(t,j)$, $u(t,j)$
- Model variables: N_C, N_K, $\mathfrak{r}$, Θ
- Numerical settings: $\mathfrak{r}$, $N_{\text{chain}}^{\text{RWM}}$, β
- Optional: $N_{\text{basis}}^{\text{FEM}}$

output:

- Global optimizer $\Gamma^*(:,j)$

Choose or generate an initial value $\Gamma^{[0]}$ (e.g., uniform initial distribution)
for $r = 1 : N_{chain}^{RWM}$ **do**
  Generate $\eta \sim \mathcal{N}(\mathbf{0},\mathbf{1})$ *and* $[0,1]$*-valued random variable* X *following a uniform distribution.*
  Compute $\Gamma' = \Gamma^{[r-1]} + \eta$ *subject to constraints (3.29),(3.30).*
  Determine acceptance rate $\mathfrak{a}(\Gamma^{[r-1]}(:,j), \Gamma'(:,j))$ for β (see (3.55))
  if $X \leq \mathfrak{a}(\Gamma^{[r-1]}(:,j), \Gamma'(:,j))$ **then**
    set $\Gamma^{[r]}(:,j) := \Gamma'(:,j)$
  else
    set $\Gamma^{[r]}(:,j) := \Gamma^{[r-1]}(:,j)$
Return $\Gamma^{[N_{\text{chain}}^{\text{RWM}}]}(:,j)$

Due to the fact that the constraints restricting parameter Θ depend on the arbitrary model f, it is in general not possible to employ the proposed stochastic optimization for the minimization of $\mathbf{L}(\Gamma,\cdot)$ or $\mathbf{L}^{\mathfrak{r}}(\Gamma,\cdot)$ with respect to Θ for fixed Γ, i.e., for Step 2 Algorithm 1. Nevertheless, in cases where the local model parameters θ_k have no or only a few restricting conditions (e.g., for model f^{kmeans} defined in (3.2)), the proposed inverse problem can be completely solved with a stochastic optimization approach. This usually has a positive effect on the memory and the computation time and simultaneously helps to avoid any locality problems. Summarizing, it is sensible to check for each choice of model whether the MCMC ansatz might be beneficial.

Further, it is important to mention that the performance of the introduced framework crucially depends on the chain's convergence to the considered distribution. Consequently, a considerable effort has been put into the development of convergence diagnostics that can be deployed to improve the results of the Metropolis algorithm [93].

In pursuance of obtaining qualitative approximations of Γ^*, a particular scheme has been proposed for the MCMC-based optimization of a Tikhonov-regularized average clustering functional [30, 37]. Essentially, this adaptive scheme is based on a very popular and recently theoretically verified convergence diagnostic [94]. It states that the optimal (with respect to the results) percentage of accepted samples of the Metropolis algorithm is 23.4% for a certain family of densities in the class of random walk Metropolis algorithms[8]. Thus, in order to improve the results, an adaptive tuning procedure is employed to keep the total number of accepted samples in the mentioned range. The details are discussed in the following paragraph and in Appendix B.

ADAPTIVE SIMULATED ANNEALING As already mentioned, the probability of any sample following the considered Boltzmann distribution $\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma)$ to be a global minimum $\Gamma^*(:,j)$ of $\mathbf{L}_j^{\mathrm{r}}(\cdot,\Theta)$ (for fixed Θ and $j \in \{1,\ldots,N_J\}$) is one if the temperature is tending to zero, i.e., the inverse temperature variable β is approaching infinity. Thus, classically a specific *cooling schedule*

$$\left\{\beta^{[1]},\beta^{[2]},\ldots,\beta^{[N_{\text{chain}}^{\text{RWM}}]}\right\} \tag{3.62}$$

is proposed priorly to the Metropolis algorithm run in order to regulate the decrease of the temperature (respectively the increase of β) during the RWM run at a certain iteration step, i.e., at a specific length of the already generated chain.

The constitution of the schedule, especially concerning the pace used to increase the value of β, is vastly important to obtain a global instead of a local minimum. It has been shown that it is possible to suffice the general environment needed to obtain a global minimum if the temperature is cooled down very slowly (for details see [40]).

However, such cooling schedules induce very long computing times and are, therefore, often not applicable to real life problems. Moreover, a fixed schedule does not adapt to random phenomena occurring during the run. More specifically, wether a proposed sample $\Gamma'(:,j)$ with minimally smaller energy value than the previous chain member $\Gamma^{[r-1]}(:,j)$, i.e.,

$$\mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma^{[r-1]}(:,j)) - \mathcal{F}_{\mathbf{L}^{\mathrm{r}},\beta}(\Gamma'(:,j)) \tag{3.63}$$

8 Although the importance of these significant three digits is obvious, they should be handled with care when the density is different from the family of densities proposed in [94], for more information see [90].

is small, is going to be accepted, depends on the magnitude of β.

For instance, for a small inverse temperature value the probability for $\Gamma'(:, j)$ to be accepted is high. However, due to the exponential nature of the Boltzmann distribution, bigger β values cause the probability for $\Gamma'(:, j)$ to become a member of the chain to approach zero. This property of the Boltzmann distribution is only desirable in a later state of the generation procedure.

Yet, in pursuance of finding a global rather than a local optimum, it is important to traverse the entire sample space. Thus, the temperature has to be increased slowly or not at all for some time at the beginning of the run in order to avoid being stuck around a local minimum.

Since the evolution of a run can not be predicted, due to the many random components, a fixed schedule does not allow enough flexibility to deal with randomly occurring local optima. This problem is addressed via an adaptive cooling schedule.

In detail this means that the variable β is changed dependent on the percentage of accepted samples $\Gamma'(:, j)$ with a smaller energy value than the current chain member, i.e.,

$$\mathcal{F}_{\mathbf{L}^r,\beta}(\Gamma'(:, j)) \leq \mathcal{F}_{\mathbf{L}^r,\beta}(\Gamma^{\lceil r-1 \rceil}(:, j)). \tag{3.64}$$

Thus, a counter is added to Algorithm 2, and the percentages are considered in a sensible frequency which might result in a change of the current inverse temperature value.

The frequency, the percentages, and the magnitude of tuning the present values should depend on the application itself. Thus, a certain level of experimental tuning is necessary to find the right setting. A complementary discussion including a pseudocode of the procedure can be found in Appendix B.

The proposed adaptive simulated annealing scheme has been tested on synthetic as well as real data sets in the context of solving clustering problems similar to the one given in (3.45) and produced promising results [30]. In general, it can be said that a suitable adaptive annealing approach is preferable to a fixed set of variables as it commonly leads to much better results.

Note that there are several alternative adaptive simulated annealing approaches that have been developed in the context of MCMC algorithms in pursuance of improving the drawbacks of a classical ansatz (see (3.62)), e.g.,

methodologies like simulated sinstering [72], simulated tempering [77], and sequential Monte Carlo [32].

Another variable that also needs to be frequently tuned or chosen carefully is the variance of the proposal density as the results vastly differ for different values. As a proposed sample $\Gamma'(:,j)$ is a noisy version of the previous chain member (see (3.56)), the energy difference only depends on the noise value $\eta(t,j)$. Thus, a big variance results in a bigger difference, causing the sample to be rejected more frequently. On the other hand, greater variance values allow to ensure that the sample space is explored more widely.

Concluding, it is necessary to have a good balance. In order to tune the noise directly, a noise factor $\mathfrak{n}$ is introduced and the original random walk proposal (see (3.56)) is slightly modified to include the new variable, i.e.,

$$\Gamma'(:,j) = \Gamma^{[r-1]}(:,j) + \mathfrak{n}\eta(t,j). \tag{3.65}$$

A detailed examination of the effect of small parameter changes on the energy value of the samples has been executed in [30, 37]. The experiments were consistent with popular convergence diagnostics.

The key idea is to try to keep the acceptance-rejection ratio, i.e., the number of accepted samples relative to the total number of rejected ones, around the verified optimal value of 23.4%. This can be achieved by regularly tuning the value of the variance $\mathfrak{n}$, which has, as established, a vast influence on the number of accepted samples.

The adaptive change of the variance factor $\mathfrak{n}$ is included as an additional feature of the proposed adaptive simulated annealing scheme. A detailed discussion on the tuning procedure can be found in Appendix B.

Summarizing, in pursuance of solving a sub optimization of a considered inverse problem, the presented RWM algorithm approach with adaptive simulated annealing scheme is a good alternative to standard minimization techniques. The benefits of the framework become especially pronounced for data with a long time-wise component and not very persistent behavior.

More specifically, the Metropolis algorithm ansatz for the Tikhonov-regularized optimization problem of finding an optimal $\Gamma^*(:,j)$ for fixed Θ and $j \in \{1,\dots,N_J\}$ has a linear computational complexity, i.e.,

$$\mathcal{O}(N_K N_{\text{basis}}^{FEM} N_{\text{chain}}^{\text{RWM}}), \tag{3.66}$$

where the standard approach with quadratic optimization is NP-complete [115].

3.6 MODEL SELECTION

The last step of the parametrization procedure is to select an *optimal* model in the set of different candidates

$$\mathcal{M}^f(N_K, N_C, N_M, u(t,j)), \tag{3.67}$$

characterized by the model parameters $P(u(t,j))$ and $\Gamma(t,j)$, dependent on the number of local regimes N_K, the maximal allowed number of transitions N_C, the memory-depth N_M, and the choice of the vector of explicit external factors $u(t,j)$.

The goal is to choose a model that has a high accuracy and, at the same time, the smallest possible number of free parameters. Thus, an optimal model should have enough explanatory power while being as simple as possible. This principle is also known as *Occam's razor* [2].

In order to determine, which model is optimal in the sense of Occam's razor, *Akaike's Information Criterion* (AIC) [3] is considered. The conceptual idea of an information criterion is to rank the different models according to the balance between their number of free parameters and their approximation quality with respect to the given data. Thus, the value of AIC can be considered to be a measurement of the disproportion between the quality of the considered model $\mathcal{M}^f(N_K, N_C, N_M, u(t,j))$ and the total number of parameters involved in the calculation of the model.

In the context of an information criterion, the quality of a model is generally given by its likelihood function. Unfortunately, it is usually not possible to directly link a likelihood function to a model of the considered nature. Yet, a modified version of AIC, applicable for the considered model framework, has been proposed in [84]. Essentially, the conformity of a model $\mathcal{M}^f(N_K, N_C, N_M, u(t,j))$ is associated with the likelihood $\mathcal{L}(N_K, N_C, N_M, u(t,j))$ of a set of parametric polynomial (conditional) probability density functions fitted to the residual processes of the local model parameters. Thus, it is necessary to assume that the scalar process of model distances follows an independent set of parametric (conditional) probability density functions

$$\phi_k(\cdot, \ldots, \cdot | N_{\phi_k}). \tag{3.68}$$

Then the likelihood, based on the model distance functional, can be defined as

$$\begin{aligned}&\mathcal{L}(N_K, N_C, N_M, u(t,j)) \\ &\quad = \prod_{j=1}^{N_J}\prod_{t=1}^{N_T}\sum_{k=1}^{N_K}\gamma_k(t,j)\phi_k\Big(g(\pi(t+1,j),\dots,\pi(t-N_M,j),\theta_k(u(t,j)))|N_{\phi_k}\Big).\end{aligned} \tag{3.69}$$

The corresponding modified version of Akaike's Information Criterion (mAIC) is defined as

$$\begin{aligned}&\mathbf{mAIC}(N_K, N_C, N_M, u(t,j), f) \\ &\qquad := -2\log(\mathcal{L}(N_K, N_C, N_M, u(t,j))) + 2|\mathcal{M}^f(N_K, N_C, N_M, u(t,j))|,\end{aligned} \tag{3.70}$$

where

$$|\mathcal{M}^f(N_K, N_C, N_M, u(t,j))| \tag{3.71}$$

denotes the total number of involved parameters. For instance, the total number of free parameters, considering the Markov model given in (3.22), sums up to

$$\begin{aligned}&|\mathcal{M}^{\text{Markov}}(N_K, N_C, N_M, u(t,j))| \\ &\qquad = \begin{cases} |\Gamma| + N_K N_S(N_S-1)(N_E+1) + |\Lambda| & \text{for } N_M = 1, \\ |\Gamma| + N_K(N_S-1)^2(N_E+1) + |\Lambda| & \text{for } N_M = 0, \end{cases}\end{aligned} \tag{3.72}$$

or, considering the logistic model defined in (3.3), is given by

$$|\mathcal{M}^{\text{logit}}(N_K, N_C, N_M, u(t,j))| = |\Gamma| + N_K(N_E + 1 + N_M) + |\Lambda|. \tag{3.73}$$

The required number $|\Gamma|$ of free variables necessary to reconstruct Γ is computed as presented in Algorithm 3. Essentially, every time-wise switch between the local regimes is considered to be a free parameter. The sum over all locations then results in the considered $|\Gamma|$.

Note that $|\Gamma|$ is less then $N_T N_J N_K$ and might even be much smaller than $N_C N_J N_K$. In other words, only the required memory, necessary to reconstruct Γ and not the full size of Γ, is counted in the total number of free parameters. In particular, this is necessary, as the optimal model with respect to mAIC would be underfitting the data for an enormous penalty term such as the one stemming from the full size of Γ.

In fact, although the regime assigning process $\Gamma(t,j)$ is explicitly dependent on the location j, it is often the case that many locations have similar

structured affiliations. Consequently, taking these similarities into account, an even smaller number of free parameters might be necessary to reconstruct Γ. Summarizing, it is important to be aware that the penalty term has to be considered with care to avoid overfitting as well as underfitting models.

Algorithm 3: Computation of $|\Gamma|$

```
input  : Γ
output: |Γ|
|Γ| = 0
for j = 1 : N_J do
    for j = 1 : N_T do
        if |Γ(t − 1, j) − Γ(t, j))| > Machine epsilon then
            |Γ| = |Γ| + 1
```

Although the introduced modified information criterion was primarily introduced to determine the optimal values for variables N_K and N_C, it has the additional advantage that it can be deployed to identify the statistical optimal model with respect to prior assumptions.

For instance, if one Markov model with and one without memory (i.e., $N_M = 1$ respectively $N_M = 0$) are fitted to the same observational data, the mAIC values determined with the proposed model-discrimination procedure can be compared to find an optimal direct mathematical model f.

In cases where the dimension of the data is relatively small with respect to the total number of parameters, required to describe the considered process best, the values computed via the mAIC are biased. In fact, this issue is a form of overfitting and a problem concerning all standard information criteria. Different approaches to address this bias have been proposed, e.g, *corrected* AIC (AICc) [17] or *improved* AIC (AICi) [8].

An alternative mechanism, often employed in the context of model validation, is the so called cross validation [19]. The principle idea is to estimate the predictive skills of each model by comparing its out-of-sample approximation to the actual data. Unfortunately, this alternative option is not generally computationally feasible for the considered non-stationary, non-homogenous models due to the high number of different necessary combinations required for the computation of a corresponding Γ^*. Details of this problem will be discussed by means of the arctic sea ice application considered in Chapter 5.

Following the idea of AICc, the potential bias of the mAIC is leveled out via an additional penalty term. Essentially, the conceptual idea is to incorporate the ratio of the number of data points to the number of parameters, i.e.,

$$\begin{aligned}&\mathbf{mAICc}(N_K, N_C, N_M, u(t,j), f) \\ &\quad := \mathbf{mAIC}(N_K, N_C, N_M, u(t,j), f) \\ &\qquad + \frac{2|\mathcal{M}^f(N_K, N_C, N_M, u(t,j))| \cdot (|\mathcal{M}^f(N_K, N_C, N_M, u(t,j))| + 1)}{N_T N_J - |\mathcal{M}^f(N_K, N_C)| - 1}.\end{aligned} \tag{3.74}$$

After inferring an optimal model with respect to the proposed mAICc, the aim is to use the model to approximate future state probabilities. The details of such approximations are discussed in the following section.

3.7 SELF-CONTAINING PREDICTIVE MODELS

Under the assumption that the global optimal model parameters $\Gamma^*(t,j)$ and $\Theta^*(u(t,j))$, minimizing the inverse problem $\mathbf{L}(\Gamma(t,j), \Theta(u(t,j)))$ given in (3.31), can be determined with the subspace algorithm (see Algorithm 1), the observed time series $\pi(t,j)$ can be approximated using the formal definition of the direct model function f, i.e.,

$$\pi(t+1,j) \approx f\left(\pi(t,j), \ldots, \pi(t-N_M,j), \sum_{k=1}^{N_K} \gamma_k^*(t,j)\theta_k^*(u(t,j))\right). \tag{3.75}$$

It is important to stress that f needs to be linear in its parameters and the model distance functional g has to be strictly convex to ensure that the approximation in (3.75) holds (details are can be found in [84]). Note that the regarded Markov model, given in (3.22), and the corresponding model distance function g, induced by the Euclidean norm (see (3.23)), have the required properties.

In order to make statements about future developments of real life processes, the prediction $\hat{\pi}(N_T+1,j)$ of the probability distribution $\pi(N_T+1,j)$ is considered. However, a direct prediction of $\pi(N_T+1,j)$ is hampered by the non-stationarity and the non-homogeneity of the model formulation. In particular, any computation to approximate $\pi(N_T+1,j)$ involves the affiliation process Γ^*, which is unknown for t greater than N_T. Thus, due to its explicit dependence on time and space, it is necessary to predict $\Gamma^*(N_T+1,j)$ before being able to compute the state probabilities outside of the considered time interval $[1, N_T]$.

Following [84], it is possible to regard the determined optimal parameter $\Gamma^*(t,j)$ as a probability distribution of a discrete process taking values in the finite set $\{1,\ldots,N_K^*\}$.

Then the problem of identifying the underlying dynamics, corresponding to $\Gamma^*(t,j)$, can be phrased as a Markov model inverse problem such as given in (3.34). This *self-contained strategy* allows to compute a probability distribution of future affiliations

$$\hat{\Gamma}(N_T+N_{\text{pred}},j)=\Gamma^*(N_T,j)\prod_{\tau=0}^{N_{\text{pred}}-1}\left(\left[P_0^\Gamma+\sum_{e=1}^{N_E}P_e^\Gamma u_e(N_T+\tau,j)\right]\right). \quad (3.76)$$

It is important to note that the model describing Γ^* is assumed to be stationary and homogenous (i.e., $N_K^\Gamma=1$). This restriction is necessary in order to evade being confronted with the problem of having to determine future states of parameters explicitly dependent on time and space again. Nevertheless, it is common to assume stationarity as well as homogeneity in the field of time series analysis.

Alternatively, the model underlying the distribution of affiliations $\Gamma^*(t,j)$ can be characterized employing multivariate logistic regression [80] (see introduction in Section 1.3). The key strength of such an approach is the non-linear structure of the associated direct model function.

Both of these stationary, homogenous models allow to estimate the required affiliations $\Gamma^*(N_T+1,j)$. The following pseudocode describes the necessary computation steps to determine a prediction $\hat{\Gamma}(N_T+1,j)$ of the regime probabilities of future assignments for a logistic or a Markov model.

Note that the number of different states is equal to the former optimal number of regimes, i.e., $N_S^\Gamma=N_K^*$ and that the number of regimes corresponding to the model characterizing Γ^* is set to be one, i.e., $N_K^\Gamma=1$, due to the assumed stationarity and homogeneity of the underlying affiliation process. Consequently, it is possible to approximate $\pi(N_T+1,j)$ via the formula given in (3.75). The details of the estimation procedure, corresponding to a Markov model, are discussed in Chapter 4 in Algorithms 6 and 7.

Algorithm 4: Self-containing prediction scheme

input :

- $\Gamma^*(t,j)$ for $t \in \{1,\dots,N_T\}$
- $u(t,j)$ for $t \in \{1,\dots,N_T+1\}$
- Model function f

output:

- $\hat{\Gamma}(N_T+1,j)$

if $f = \mathit{Markov}$ **then**
  Determine
  $\mathbf{L}(P^{\Gamma}(u(t,j))) = \sum_{j=1}^{N_J}\sum_{t=1}^{N_T} \|\Gamma^*(t+1,j) - \Gamma^*(t,j)P^{\Gamma}(u(t,j))\|_2^2 \to \min_{P^{\Gamma}(u(t,j))}$
  via stationary, homogenous Markov regression.
  for $j = 1 : N_J$ **do**
    $\hat{\Gamma}(N_T+1,j) = \Gamma^*(N_T,j)P^{\Gamma}(u(N_T,j))$
if $f = \mathit{logit}$ **then**
  Infer
  $\mathbf{L}(B^{\Gamma}) = \sum_{t=1}^{N_T}\sum_{j=1}^{N_J} \|\Gamma^*(t,j) - \theta^{\text{logit}}(B^{\Gamma},u(t,j))\|_2^2 \to \min_{B^{\Gamma}}$
  via multivariate logistic regression.
  for $j = 1 : N_J$ **do**
    $\hat{\Gamma}(N_T+1,j) = \theta^{\text{logit}}(B^{\Gamma},u(N_T+1,j))$

4

TEST MODEL SYSTEMS

Before employing the introduced non-stationary and non-homogenous regression framework to characterize the arctic sea ice dynamics, the capability of the method is examined on two different artificial data sets. The considered test model systems include a synthetic dynamical process

$$\sigma^{\text{syn}}(t, j, l) \tag{4.1}$$

driven by external forces, which are not made fully available for the parametrization procedure. Essentially, the ability of the model to compensate missing information is tested experimentally. In order to access the quality of the obtained model, the results of standard approaches such as ANN [9, 11, 55, 69] and SVM [27, 102] are presented as a reference.

To be able to distinguish between the various results computed with different methodologies (e.g., Markov and logit), labels are added to the inferred model parameters (e.g., $\Gamma^{\text{Markov}}(t, j)$) and the associated approximations (e.g., $\pi^{\text{Markov}}(t, j)$) of the data $\pi(t, j)$. Note that some variables, such as the total number of entries N_E in the vector of explicit external factors $u^{\text{syn}}(t, j)$, the finite number of states N_S and the number of considered time steps N_T, are universal for all techniques and, thus, are not labeled.

Further, the synthetic parameters used to generate the toy examples are tagged with the superscript *syn* (e.g., $\Gamma^{\text{syn}}(t, j)$). Additionally, a superscripted asterisk is used for optimal (in the sense that the lowest mAICc value is attained for the corresponding model) model parameters and variables.

In order to test the out-of-sample performance of the computed model, the synthetic data set is split into a *training sequence*

$$\{1, \ldots, N_{T_{\text{train}}}\}, \tag{4.2}$$

used to infer the model parameters, and a *test sequence*

$$\{N_{T_{\text{train}}} + 1, \ldots, N_T\}, \tag{4.3}$$

used to validate the model approximations.

The first toy example is chosen to have ideal experimental settings (any relevant influencing quantities are provided for the computations, i.e., there are no unresolved external factors). The aim is to show the general utility of the proposed parametrization ansatz under good conditions.

In order to explore the characteristic property of the Markov model to reflect the unavailable influences on a system in an explicit time and space dependency, most of the external factors used to generate the second example data set are assigned to be implicit (i.e., $N_E = 1$ and $N_I = 100$). In detail that means that these factors are not made available for the parametrization procedure.

The synthetic process $\sigma^{\text{syn}}(t, j, l)$ is chosen to be Markovian for both exemplary artificial dynamical systems. In fact, the associated transition matrix

$$P^{\text{syn}}(t, j, u^{\text{syn}}(t, j)) \tag{4.4}$$

is calculated by means of a weighted sum of N_K^{syn} matrices of the particular structure shown in (3.35).

The weights $\gamma_k^{\text{syn}}(t, j)$ are randomly generated. Yet, a certain level of persistency is forced on the random process via the previously chosen N_C^{syn}, restricting the number of transitions between the N_K^{syn} regimes in the considered time interval. The details of the computation of the corresponding data are explained in the following (see Algorithms 5 and 6).

The required affiliations $\gamma_k^{\text{syn}}(t, j)$ can be generated for different input values N_K^{syn}, N_C^{syn}, N_T, and N_J with Algorithm 5. Note that $\gamma_k^{\text{syn}}(t, j)$ takes values in the set $\{0, 1\}$ for reasons of simplicity and is subject to the constraints (3.29), (3.30), and (3.48). In particular, the required time-wise persistency of $\Gamma^{\text{syn}}(t, j)$ is enforced in Lines 3-12 of Algorithm 5.

As a consequence of the restriction of $\gamma_k^{\text{syn}}(t, j)$ to the set $\{0, 1\}$, the transition matrix $P^{\text{syn}}(t, j, u^{\text{syn}}(t, j))$ corresponding to the given local models $P^{k\,\text{syn}}(u^{\text{syn}}(t, j))$ can be approximated as follows:

$$P^{\text{syn}}(t, j, u^{\text{syn}}(t, j)) \approx \sum_{k=1}^{N_K^{\text{syn}}} \gamma_k^{\text{syn}}(t, j) P^{k\,\text{syn}}(u^{\text{syn}}(t, j)), \tag{4.5}$$

where $P^{k\,\mathrm{syn}}(u^{\mathrm{syn}}(t,j))$ is chosen to have the linear structure given in (3.35). The requirements for this approximation have already been outlined in Section 3.7, and a detailed derivation (for purely time-dependent model parameters) can be found in [84].

Consequently, the artificial ensemble data $\pi^{\mathrm{syn}}(t,j)$ can be calculated via randomly (probability given by the transition matrix $P^{\mathrm{syn}}(t,j,u^{\mathrm{syn}}(t,j))$) generating an ensemble of N_{ens} realizations of the process $\sigma^{\mathrm{syn}}(t,j,l)$ and determining the relative frequencies (see (2.1)). The computation of the artificial local models $P^{k\,\mathrm{syn}}(u^{\mathrm{syn}}(t,j))$, given matrices $P_e^{k\,\mathrm{syn}}$, is shown in Line 4 of Algorithm 6.

Note that the synthetic transition matrix $P^{\mathrm{syn}}(t,j,u^{\mathrm{syn}}(t,j))$, used to generate the artificial data, has a linear dependency on the implicit external factors. Essentially, the entries of $\bar{u}(t,j)$ (in particular the ones for $e > N_E$) are treated as explicit external factors for the generation of the data, i.e., a set of local model matrices

$$P_{N_E+1}^{k\,\mathrm{syn}}, \ldots, P_{N_E+N_I}^{k\,\mathrm{syn}}, \tag{4.6}$$

associated with the implicit external factors in $\bar{u}^{\mathrm{syn}}(t,j)$, is chosen for $k \in \{1,\ldots,N_K^{syn}\}$. During the parametrization procedure, however, these unresolved entries of the vector $\bar{u}^{\mathrm{syn}}(t,j)$ are assigned to be unknown/unavailable.

Summarizing, a set N_F of synthetic model matrices is chosen and the transition matrix $P^{\mathrm{syn}}(t,j,u^{\mathrm{syn}}(t,j))$ is calculated using the assumed model structure, given in (3.35), and Equation (4.5) (see Line 4 of Algorithm 6).

To generate the corresponding ensemble data, the relative frequency of N_{ens} different realizations of the artificial process in each cell j for a fixed t (see Lines 5-10 of Algorithm 6) is determined.

For the required sampling step in Lines 7-8 of Algorithm 6, it possible to use standard methodologies such as rejection sampling (also known as the acceptance-rejection method) [22, 92, 116].

4.1 TOY EXAMPLE 1: IDEAL CONDITIONS

Here the focus is on the basic attributes of the proposed framework and especially on its feasibility under ideal conditions. More specifically, that means that all the external factors, comprised in the vector $\bar{u}^{\mathrm{syn}}(t,j)$, used to generate the considered artificial dynamical process $\sigma^{\mathrm{syn}}(t,j,l)$ are available for the model inference. Yet, the underlying model of the process is chosen to

Algorithm 5: Generate synthetic affiliation $\Gamma^{\text{syn}}(t, j)$

input :

- N_K^{syn} synthetical number of local models
- N_C^{syn} synthetical maximal number of transitions
- Spatial and time dimension N_T and N_J

output:

- $\Gamma^{\text{syn}}(t, j)$ synthetically generated affiliations

for $j = 1 : N_J$ **do**
  $\gamma_k^{\text{syn}}(:, j) = [\,] \quad \forall\, k \in \{1, \ldots, N_K\}$
  for $c = 1 : N_C^{syn}$ **do**
    $N_{\text{dummy}} = \textbf{round}\Big(2N_T / (N_C^{\text{syn}} \cdot \textbf{rand}([0, 1]))\Big)$
    $\text{dummy0} = (0, \ldots, 0) \in \mathbb{R}^{1 \times N_{\text{dummy}}}$
    $\text{dummy1} = (1, \ldots, 1) \in \mathbb{R}^{1 \times N_{\text{dummy}}}$
    $r = \textbf{rand}(\{1, \ldots, N_K^{\text{syn}}\})$
    for $k = 1 : N_K^{syn}$ **do**
      if $r == k$ **then**
        $\gamma_k^{\text{syn}}(:, j) = [\gamma_k^{\text{syn}}(:, j)\ \text{dummy1}]$
      else
        $\gamma_k^{\text{syn}}(:, j) = [\gamma_k^{\text{syn}}(:, j)\ \text{dummy0}]$
  if $\textbf{length}(\gamma_1^{syn}(:, j)) \geq N_T$ **then**
    $\gamma_k^{\text{syn}}(:, j) = \gamma_k^{\text{syn}}(1 : N_T, j) \quad \forall\, k \in \{1, \ldots, N_K^{\text{syn}}\}$
  else
    $N_{\text{dummy}} = N_T - \textbf{length}(\gamma_1^{\text{syn}}(:, j))$
    $\text{dummy0} = (0, \ldots, 0) \in \mathbb{R}^{1 \times N_{\text{dummy}}}$
    $\text{dummy1} = (1, \ldots, 1) \in \mathbb{R}^{1 \times N_{\text{dummy}}}$
    $\gamma_1^{\text{syn}}(:, j) = [\gamma_1^{\text{syn}}(:, j)\ \text{dummy1}]$
    $\gamma_k^{\text{syn}}(:, j) = [\gamma_k^{\text{syn}}(:, j)\ \text{dummy0}] \quad \forall\, k \in \{2, \ldots, N_K^{\text{syn}}\}$
  $\Gamma^{\text{syn}}(:, j) = [\gamma_1^{\text{syn}}(:, j), \ldots, \gamma_{N_K^{\text{syn}}}^{\text{syn}}(:, j)]$

be non-stationary and non-homogenous and, consequently, is more complex than assumed in many standard models of time series analysis.

As the second artificial parametrization problem presented in Section 4.2 is posed under considerably worse conditions (with respect to the available external influences), the first example also serves as a reference for the general approximative abilities of the methodology under better conditions.

The synthetic data, associated with $\sigma^{\text{syn}}(t, j, l)$, is generated with Algorithms 5 and 6 for the following input values: $N_C^{\text{syn}} = 10$, $N_K^{\text{syn}} = 2$, $N_T = 400$,

Algorithm 6: Generate synthetic data $\pi^{\text{syn}}(t,j)$

input :

- $\Gamma^{\text{syn}}(t,j)\ \forall\ t$ and j (see Algorithm 5) with corresponding N_K^{syn}, N_T, and N_J
- Ensemble size N_{ens} and finite set of discrete states $\{s_1,\dots,s_{N_S}\}$
- Vector of external factors $\bar{u}^{\text{syn}}(t,j) \in \mathbb{R}^{N_F\times 1}$
- Model matrices $P_0^{k\ \text{syn}},\dots,P_{N_E}^{k\ \text{syn}},P_{N_E+1}^{k\ \text{syn}},\dots,P_{N_E+N_I}^{k\ \text{syn}}$ with $k \in \{1,\dots,N_K^{\text{syn}}\}$

output:

- Synthetic ensemble data $\pi^{\text{syn}}(t,j)$ associated with artificial process $\sigma^{\text{syn}}(t,j,l)$

Initialize $\sigma^{\text{syn}}(0,j,l) = \textbf{rand}\{s_1,\dots,s_{N_S}\}\ \forall\ j \in \{1,\dots,N_J\}$, $l \in \{1,\dots,N_{\text{ens}}\}$
for $t = 1 : N_T$ **do**
  for $j = 1 : N_J$ **do**
    $P^{\text{syn}}(t,j,\bar{u}(t,j)) = \sum_{k=1}^{N_K} \gamma_k(t,j)\left(P_0^{k\ \text{syn}} + \sum_{e=1}^{N_F} P_e^{k\ \text{syn}}\bar{u}_e^{\text{syn}}(t,j)\right)$
    for $l = 1 : N_{ens}$ **do**
      $h = \textbf{index}(\sigma^{\text{syn}}(t-1,j,l))$
      $\sigma^{\text{syn}}(t,j,l) =$
      $\begin{cases} s_1 & \text{with probability } \{P^{\text{syn}}(t,j,\bar{u}^{\text{syn}}(t,j))\}_{h1} \\ & \vdots \\ s_{N_S} & \text{with probability } \{P^{\text{syn}}(t,j,\bar{u}^{\text{syn}}(t,j))\}_{hN_S} \end{cases}$
      (see rejection sampling [22, 92, 116])
    for $i = 1 : N_S$ **do**
      $\pi_i^{\text{syn}}(t,j) = \textbf{counter}(\sigma^{\text{syn}}(t,j,l) = s_i)/N_{\text{ens}}$

$N_J = 24$, $N_S = 2$, $N_E = 2$, and $N_I = 0$. Further, the external influence vector $\bar{u}^{\text{syn}}(t,j)$ consists of entries

$$\bar{u}_1^{\text{syn}}(t,j) := \sin^2\left(\frac{4\pi t}{360} + \frac{j}{20}\right) \tag{4.7}$$

and

$$\bar{u}_2^{\text{syn}}(t,j) := \underset{\text{r}\in\ \text{neigh}(j)}{\textbf{average}}(\pi(t-1,r)), \tag{4.8}$$

where the second entry represents existing spatial correlations between direct neighboring cells. Note that the locations j in this example are associated

with cells on a honeycomb lattice. This entails that each location j has six neighbors, all sharing an edge with the respective cell.

The model matrices, used to synthetically generate the data, are defined as follows:

$$P_0^{1\,\mathrm{syn}} = \begin{bmatrix} 0.7 & 0.3 \\ 0.7 & 0.3 \end{bmatrix}, P_1^{1\,\mathrm{syn}} = \begin{bmatrix} 0.28 & -0.28 \\ 0.28 & -0.28 \end{bmatrix}, P_2^{1\,\mathrm{syn}} = \begin{bmatrix} -0.01 & 0.01 \\ -0.01 & 0.01 \end{bmatrix} \tag{4.9}$$

and

$$P_0^{2\,\mathrm{syn}} = \begin{bmatrix} 0.3 & 0.7 \\ 0.3 & 0.7 \end{bmatrix}, P_1^{2\,\mathrm{syn}} = \begin{bmatrix} 0.24 & -0.24 \\ 0.24 & -0.24 \end{bmatrix}, P_2^{2\,\mathrm{syn}} = \begin{bmatrix} 0.05 & -0.05 \\ 0.05 & -0.05 \end{bmatrix}. \tag{4.10}$$

In order to test whether it is possible to infer a good model description of the considered synthetical process, the proposed regression framework is applied to the training set, i.e., $\pi(t, j)$ with $t \in \{1, \ldots, 360\}$ ($N_{T_{\mathrm{train}}} = 360$) for different model assumptions. Essentially, that means to compute a variation of models

$$\mathcal{M}^f(N_K, N_C, N_M, u(t, j)) \tag{4.11}$$

for $N_C \in \{3, 5, 7, 10, 15, 20\}$, $N_K \in \{1, 2, 3\}$, $N_M \in \{0, 1\}$, and different direct mathematical model functions, i.e., f^{logic} and f^{Markov}. The required numerical settings of all runs are: $N_{\mathrm{anneal}}^{\mathrm{FEM}} = 10$ and $N_{\mathrm{tol}}^{\mathrm{FEM}} = 0.0000000001$.

As outlined in Section 3.6, the selection of an appropriate or optimal model in the respective set of different models is realized via the mAICc. Figures 6 and 7 show the mAICc values that are attained for each of the computed models. The values can all be regarded on the same scale, yet, the significant differences between the results of the Markov and the logistic models are obvious.

This result is in line with the expectations, as a Markov model was used to generate the artificial data. The lowest mAICc value is attained for the model $\mathcal{M}^{\mathrm{Markov}}(2, 10, 0, \bar{u}^{\mathrm{syn}}(t, j))$. Concluding, the artificially chosen variables N_C^{syn} and N_K^{syn} are correctly identified.

In this artificial setting it is also possible to compare the determined model matrices to the matrices $P_0^{k\,\mathrm{syn}}, \ldots, P_{N_E}^{k\,\mathrm{syn}}$ with $k \in \{1, \ldots, N_K^{\mathrm{syn}}\}$ (see (4.9) and (4.10)), used to generate the synthetic data:

$$\begin{aligned} P_0^{1\,\mathrm{Markov}} &= \begin{bmatrix} 0.6999 & 0.3001 \\ 0.3001 & 0.6999 \end{bmatrix}, P_1^{1\,\mathrm{Markov}} = \begin{bmatrix} 0.2801 & -0.2801 \\ 0.2801 & -0.2801 \end{bmatrix}, \\ P_2^{1\,\mathrm{Markov}} &= \begin{bmatrix} -0.0125 & -0.0515 \\ -0.0125 & -0.0515 \end{bmatrix} \end{aligned} \tag{4.12}$$

and

$$P_0^{2\,\text{Markov}} = \begin{bmatrix} 0.3003 & 0.69971 \\ 0.3003 & 0.6997 \end{bmatrix}, P_1^{2\,\text{Markov}} = \begin{bmatrix} 0.24 & -0.24 \\ 0.24 & -0.24 \end{bmatrix}, \\ P_2^{2\,\text{Markov}} = \begin{bmatrix} 0.0515 & -0.0515 \\ 0.0515 & -0.0515 \end{bmatrix}. \tag{4.13}$$

Supplementary to the parametrization with the non-stationary, non-homogenous regression, data approximations are determined via trained ANNs [9, 69, 55, 11] and SVMs [27, 102].

This additional comparison opportunity allows to consider the results of the proposed regression technique in the context of the model inference quality of two standardly employed data analysis methods. A brief description of these two pattern recognition techniques is given in Chapter 1. In the context of ANNs, transfer functions $\Psi(t)$ as well as a network structure have to be chosen. As MLPs with one hidden layer and logistic transfer functions (see (1.13))[1] have been shown to be universal approximators [55], this particular type of networks is used to estimate the distribution $\pi^{\text{syn}}(t,j)$[2].

Further, different networks

$$\mathcal{N}(N_{\text{neurons}}^{\text{ANN}}) \tag{4.14}$$

for various numbers of hidden neurons, i.e.,

$$N_{\text{neurons}}^{\text{ANN}} \in \{5, 10, 15, 20, 25, 30, 40, 50\}, \tag{4.15}$$

are trained. For the training of a network, the Levenberg-Marquardt backpropagation is employed. Analogous to the additional iterations of the regression framework, which are deployed to approach a global minimum, simulated annealing steps are required to determine an optimal network, i.e., $N_{\text{anneal}}^{\text{ANN}} = 10$. A network is considered to be optimal with respect to the hidden neurons if it produces the smallest residuals, i.e.,

$$\sum_{j=1}^{N_J} \sum_{t=1}^{N_T} \left\| \pi^{\text{syn}}(t,j) - \pi^{\mathcal{N}(N_{\text{neurons}}^{\text{ANN}})}(t,j) \right\|_2^2 \rightarrow \min_{\mathcal{N}(N_{\text{neurons}}^{\text{ANN}})}. \tag{4.16}$$

1 Note that a different non-linear activation function such as the hyperbolic tangent function, given in (1.12), can be used as well. Here the logistic function $\Psi^{\text{sigmoid}}(t)$ is used for every neuron in the hidden layer.

2 Note that the total number of samples is $N_L = N_T \cdot N_J$ and that the time-wise (respectively location-wise) order of the samples is complete irrelevant for the training of the considered networks.

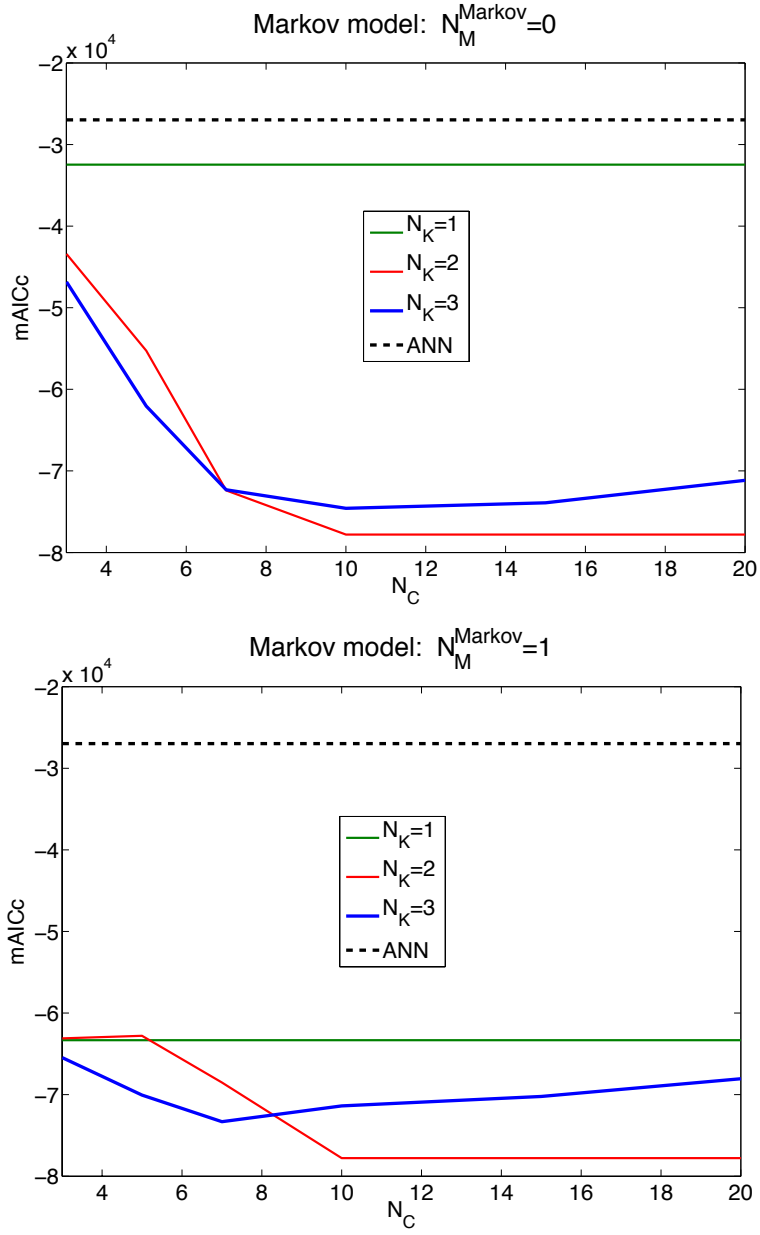

Figure 6: *The mAICc values for different Markov models* $\mathcal{M}^{Markov}(N_K, N_C, N_M, u(t,j))$ *with* $N_K \in \{1,2,3\}$, $N_C \in \{3,5,7,10,15,20\}$, *and* $N_M \in \{0,1\}$ *are shown. Additionally, the mAICc value of the optimal (with respect to mAICc) ANN model is shown.*

The smallest residuals can be obtained for a network with 20 hidden neurons, i.e., $N^{\text{ANN}*}_{\text{neurons}} = 20$. Thus, the corresponding network $\mathcal{N}(20)$ is used to compute approximations $\pi^{\mathcal{N}(20)}(t,j)$ of $\pi^{\text{syn}}(t,j)$. A comparison of approximations $\pi^{\text{Markov}}(t,j)$ and $\pi^{\mathcal{N}(20)}(t,j)$ and the data $\pi^{\text{syn}}(t,j)$ for two example locations and $t \in \{1,\ldots,400\}$ is shown in Figure 8.

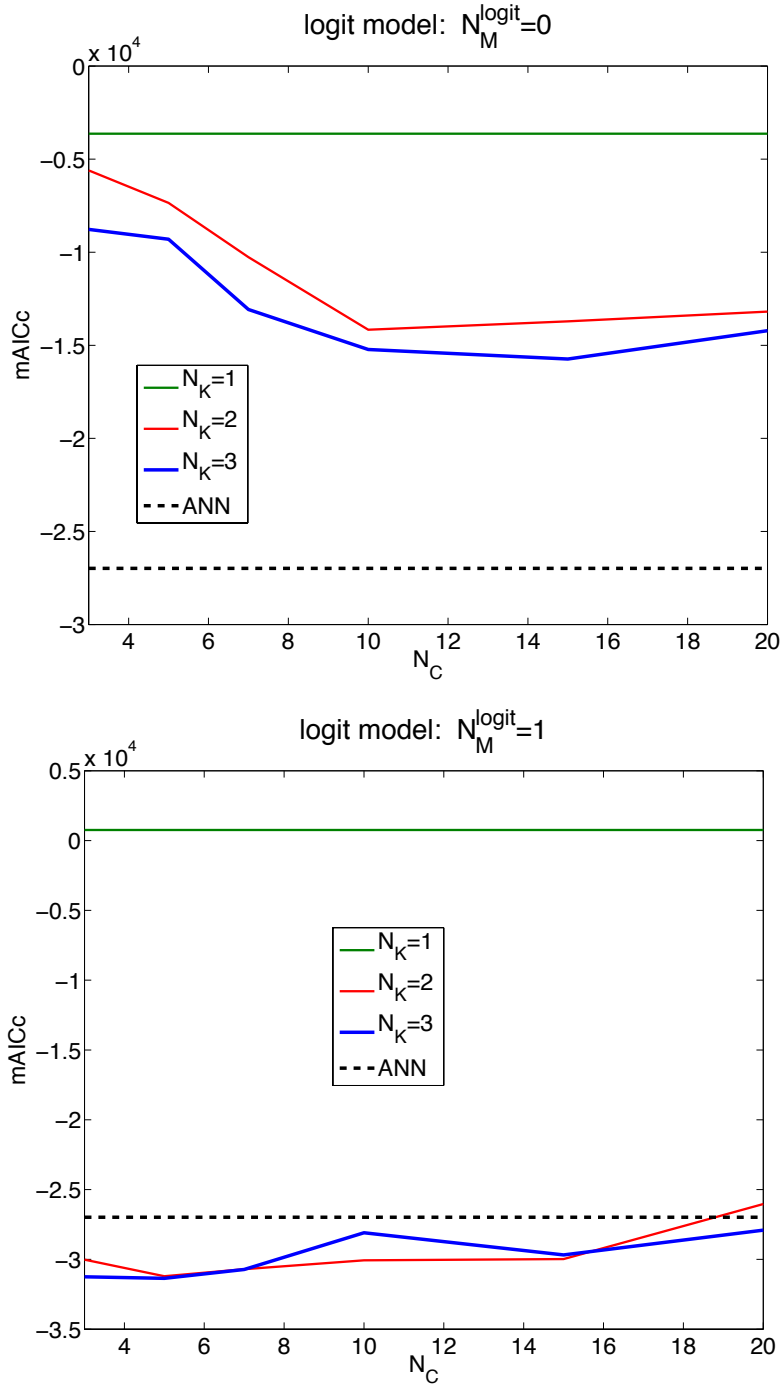

Figure 7: *The mAICc values for different logit models* $\mathcal{M}^{logit}(N_K, N_C, N_M, u(t,j))$ *for* $N_K \in \{1,2,3\}$, $N_C \in \{3,5,7,10,15,20\}$, *and* $N_M \in \{0,1\}$ *are displayed. Additionally, the mAICc value of the optimal (with respect to mAICc) ANN model is shown.*

For the training sequence, $\pi^{\text{Markov}}(t,j)$ is computed using the optimal model

$$\mathcal{M}^{\text{Markov}}(2, 10, 0, \bar{u}^{\text{syn}}(t,j)), \tag{4.17}$$

i.e., the corresponding $\Gamma^*(t,j)$ and $P^*(u(t,j))$, and employing Algorithm 6. For the test sequence, i.e., $t \in \{361,\dots,400\}$, the self-contained strategy, outlined in Section 3.7, is used to identify the model of the dynamics of the regime affiliations $\Gamma^*(t,j)$.

Different model functions f^{Markov} and f^{logit} are considered to characterize the process $\Gamma^*(t,j)$, and the lowest mAICc value is attained for a Markov model with memory (see Table 2 of mAICc values in Appendix C.1). The computational details necessary to obtain $\hat{\pi}^{\text{Markov}}(t,j)$, i.e., the out-of-sample approximations, are given in Algorithm 7.

Algorithm 7: Prediction

input :

- $P^{\Gamma}(u^{\text{syn}}(t,j))$
- Maximal prediction depth N_{pred}
- $u^{\text{syn}}(t,j)$ for $t \in \{1,\dots,N_T\}$

output:

- Prediction error $\varpi(j,\tau)$ with $\tau \in \{1,\dots,N_{\text{pred}}\}$
- $\hat{\pi}^{\text{Markov}}(t,j)$ with $t \in \{N_{T_{\text{train}}}+1,\dots,N_T\}$

1 **for** $j = 1 : N_J$ **do**
2 **for** $\tau = 1 : N_{pred}$ **do**
3 $\hat{\Gamma}(N_{T_{\text{train}}}+\tau,j) = \Gamma^*(N_{T_{\text{train}}},j) \prod_{h=0}^{\tau-1} P^{\Gamma}(u^{\text{syn}}(N_{T_{\text{train}}}+h,j))$ (see Eq. (3.76))
4 Generate $\hat{\pi}(N_{T_{\text{train}}}+\tau,j)$ employing Algorithm 6 (Lines 3 to 10) using regime probabilites $\hat{\Gamma}(N_{T_{\text{train}}}+\tau,j)$ to compute the affiliations
5 $\varpi(j,\tau) = \|\pi^{\text{syn}}(N_{T_{\text{train}}}+\tau,j) - \hat{\pi}^{\text{Markov}}(N_{T_{\text{train}}}+\tau,j)\|_2^2$

The graphs of Figure 8 show that the models are capable of qualitatively estimating the original data set for the considered dynamical system and indicates a superiority of the Markov models. In particular, the out-of-sample performance is promising. Yet, it is important to stress that the prediction error grows when the prediction depth is increased. This can be seen best considering the relative mean prediction error

$$\varpi_{\text{rel}}(\tau) = 100 \times \underset{j}{\textbf{mean}} \left(\frac{\varpi(j,\tau)}{\|\pi^{\text{syn}}(N_{T_{\text{train}}}+\tau,j)\|_2^2} \right), \tag{4.18}$$

which can be computed via Algorithm 7 for the considered data set, i.e., $N_{\text{pred}} = 23$. The error $\varpi_{\text{rel}}(\tau)$ for the considered approximations, computed with the model $\mathcal{M}^{\text{Markov}}(2, 10, 0, \bar{u}^{\text{syn}}(t, j))$ and the network $\mathcal{N}(N^{\text{ANN}}_{\mathcal{N}(20)})$, is displayed in the lower panel of Figure 9. Note that the Markov model can compete with the considered ANN in terms of the relative prediction error and even performs slightly better.

Further, the considered artificial dynamical process is modeled with SVMs. In order to infer the best fit, different kernel functions are considered. In detail, linear, quadratic, polynomial (see (1.9)) and radial basis kernel functions (see (1.10)) are used. As the artificial dynamical system is assumed to be given only in form of ensemble data and no additional data is available with respect to the actual state assignments of the process on a microscopic scale, it is necessary to set a threshold of 0.5 and to round $\pi^{\text{syn}}(t, j)$ accordingly so that the data has two categories, i.e., two classes. Consequently, it is possible to employ SVMs to characterize the dynamics underlying these state associations.

Analogous to the inference of an optimal network with respect to the number of neurons, an optimal SVM is selected by means of the residuals. The smallest values are attained for an SVM with radial basis kernel function. The results computed with the optimal SVM are shown along with the rounded approximations $\pi_1^{\text{Markov}}(t, j)$ and $\pi_1^{\mathcal{N}(20)}(t, j)$ and the artificial data $\pi_1^{\text{syn}}(t, j)$ for a fixed location in the upper panel of Figure 9.

The class affiliations determined via the employed SVM are mostly consistent with the rounded state probabilities. Occasionally wrong assignments occur due to state probabilities close to the threshold 0.5. As the nature of the data requires to determine state assignments (by rounding) to be able to employ SVMs, a different ansatz via ANN or the proposed Markov regression is more sensible.

Concluding, the proposed framework has been shown to be capable of inferring a very good approximation of the underlying model. However, it is important to emphasize that the considered synthetic process is an ideal example for the proposed framework. Consequently, the promising results always have to be considered in the context of these good conditions. In order to provide a contrast to this ideal scenario, the techniques abilities are also tested for an artificial system, where the relevant exterior factors influencing the dynamical process are not available.

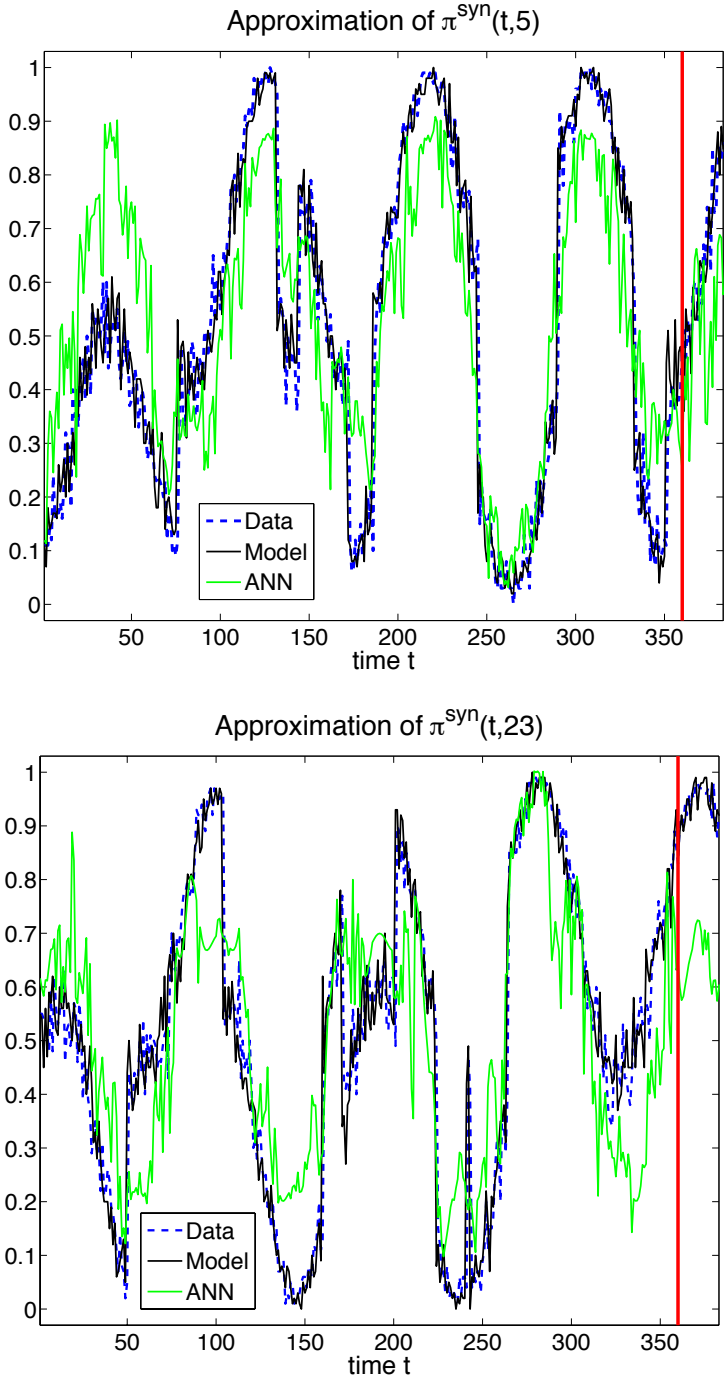

Figure 8: *Two approximations of the synthetic data $\pi_1^{syn}(t,j)$, one computed by means of model $\mathcal{M}^{Markov}(2,10,0,\bar{u}^{syn}(t,j))$ (black) and the other one determined via a network $\mathcal{N}(20)$ (green line) for two different locations $j = 5$ (see top panel) and $j = 23$ (see bottom panel), are shown. The synthetic reference values $\pi_1^{syn}(t,j)$ are also visualized. The end of the training component of the data, i.e., $t = 360$, is marked with a red vertical line.*

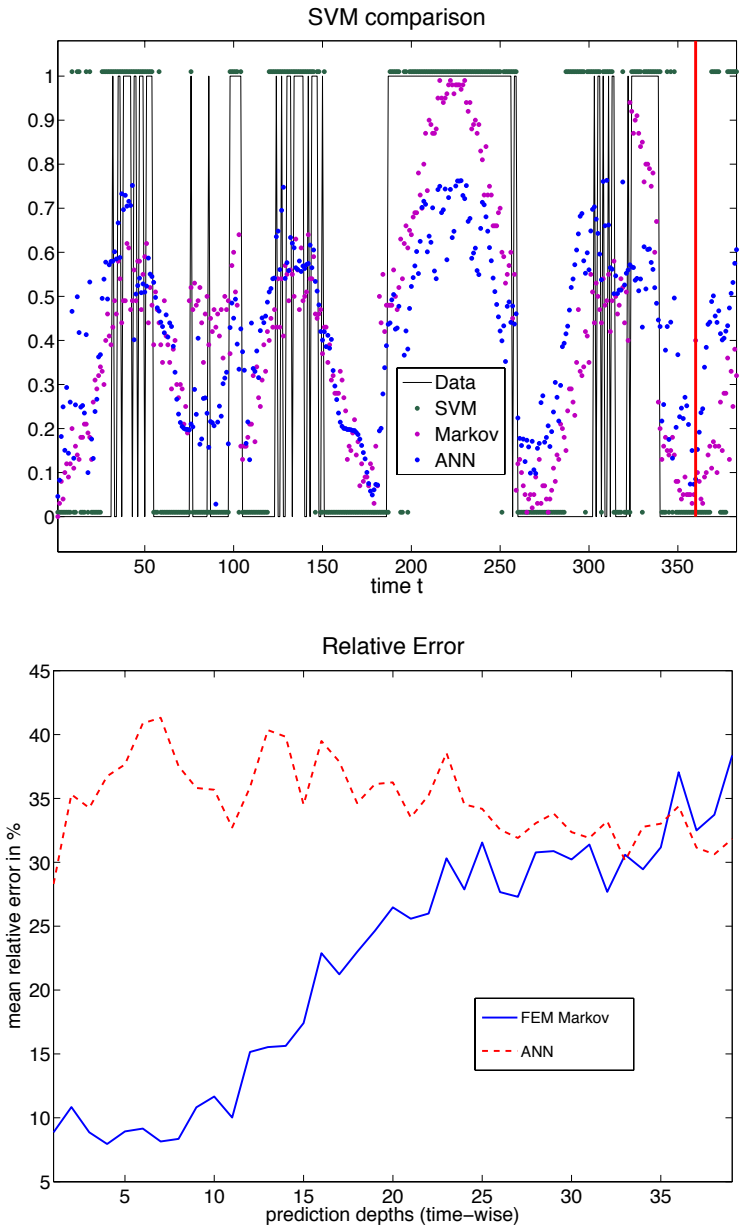

Figure 9: *The panel on the top shows the resulting SVM state assignments for* $t \in \{1, \dots, 400\}$ *(in green) and the corresponding rounded data* $\pi_1^{syn}(t, j)$ *(black). Further, the dotted approximations of the synthetic data, computed by means of the network* $\mathcal{N}(20)$ *(in blue), and the model* $\mathcal{M}^{Markov}(2, 10, 0, \bar{u}^{syn}(t, j))$ *(in pink) are presented. The mean relative error* $\omega_{rel}(\tau)$ *in % of the approximations* $\pi_1^{Markov}(t, j)$ *and* $\pi_1^{\mathcal{N}(20)}(t, j)$ *is considered in the lower panel of the figure. Note that the error depends on the prediction depth* $\tau \in \{1, \dots, 39\}$*, i.e.,* $N_{pred} = 39$*. The computational details can be found in Algorithm 7.*

4.2 TOY EXAMPLE 2: STRONG IMPLICIT INFLUENCES

As the introduced Markov model is considered for the data-based parametrization of dynamical systems driven by exterior factors, a particular emphasis

is placed upon the fact that some of the relevant exterior quantities might not be available. Consequently, the model is designed to entail the possibility to describe a joint impact of these unknown influences via an explicit dependency on time and location.

This attribute of the model is numerically tested for an artificial system $\sigma^{\text{syn}}(t,j,l)$, directly influenced by $N_F = 101$ external factors, where only approximately 1% of the exterior quantities are given for the inference of a corresponding model (i.e., $N_E = 1$ and $N_I = 100$). Summarizing, the conceptual advantage, attributed to the proposed Markov model and the corresponding non-stationary, non-homogenous regression, is numerically investigated by means of an ill-posed problem.

As already mentioned above, the artificial dynamical process is defined to have the proposed Markov structure, given in (3.22), and is binary in the sense that it takes values in the set $\{s_1, s_2\}$ (i.e., $N_S = 2$). As the aim is to examine the ability of the model to describe unknown influences via the time- and space-dependent regime affiliation $\Gamma(t,j)$, the artificial process $\sigma^{\text{syn}}(t,j,l)$ is defined to be stationary and homogenous. Essentially that means that it is defined by one regime ($N_K^{\text{syn}} = 1$) given by the model matrices:

$$P_0^{1\text{ syn}} = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, P_1^{1\text{ syn}} = \begin{bmatrix} 0.05 & -0.05 \\ 0.05 & -0.05 \end{bmatrix}, P_2^{1\text{ syn}} = \begin{bmatrix} 0.42 & -0.42 \\ 0.42 & -0.42 \end{bmatrix} \quad (4.19)$$

and

$$P_{e+2}^{1\text{ syn}} = \begin{bmatrix} 0.0002 & -0.0002 \\ 0.0002 & -0.0002 \end{bmatrix} \quad \forall\, e \in \{1, \dots N_I - 2\}. \quad (4.20)$$

The corresponding external factors are

$$\bar{u}_1^{\text{syn}}(t,j) := \underset{\text{r}\in\text{ neigh(j)}}{\textbf{average}}(\pi(t-1,r)) \quad (4.21)$$

and

$$\bar{u}_e^{\text{syn}}(t,j) = \begin{cases} \sin^2\left(\frac{2\pi te}{360} + \frac{j}{20}\right) & \textbf{rand}(e) > 0.5 \\ \cos^2\left(\frac{2\pi te}{360} + \frac{j}{20}\right) & \text{otherwise} \end{cases} \quad (4.22)$$

for $e \in \{2, \dots, N_F\}$, where (4.21) is set to be given for the parametrization, i.e., is an explicit external factor. Note that the remaining factors, defined in (4.22), are considered to be unknown quantities, i.e., implicit external factors. Due to the fact that the entries of the matrix $P_2^{1\text{ syn}}$ are comparatively large, the first implicit external quantity is the main influencing factor. Concluding, the data $\pi^{\text{syn}}(t,j)$ associated with the defined dynamical process

is generated using Algorithm 6 for a fixed $\Gamma^{\text{syn}}(t,j) := \mathbf{ones}(1, N_T, N_J)$. Again, the generated artificial data is divided (time-wise) into a training and a test sequence, and a set of non-stationary, non-homogenous models[3] is trained on $\pi^{\text{syn}}(t,j)$ for $t \in \{1,\dots,360\}$ and for all j. The parametrization procedure is executed for values $N_K \in \{1,2,3,4,5\}$, $N_C \in \{5,10,15,20,25\}$, $N_M \in \{0,1\}$, and $f \in \{\text{Markov}, \text{logit}\}$, resulting in a set of 100 different models $\mathcal{M}^f(N_K, N_C, N_M, u^{\text{syn}}(t,j))$. In this collection of potential candidates, an optimal model is selected via the proposed mAICc. The mAICc values, obtained for models

$$\mathcal{M}^{\text{Markov}}(N_K, N_C, 0, u^{\text{syn}}(t,j)) \quad \text{and} \quad \mathcal{M}^{\text{logit}}(N_K, N_C, 1, u^{\text{syn}}(t,j)) \tag{4.23}$$

for different values for N_K and N_C, are shown in the panels of Figure 10. As illustrated, the mAICc results corresponding to different logistic models with memory have higher values. Consequently, the lowest mAICc value is attained for

$$\mathcal{M}^{\text{Markov}}(4, 15, 0, u^{\text{syn}}(t,j)) \tag{4.24}$$

(see turquoise line in upper panel of Figure 10). Thus, the originally stationary and homogenous synthetic process is fitted to a model with parameters explicitly dependent on time and location. A comparison of the resulting approximation and the real data is shown in Figure 11.

Additionally, an approximation of the synthetic data computed via a network $\mathcal{N}(10)$ is shown. In order to find a qualitative network, a set of six networks, associated with different neurons, i.e.,

$$N^{\text{ANN}}_{\text{neurons}} \in \{5, 10, 15, 20, 25, 30, 40, 50\}, \tag{4.25}$$

is determined by means of the training data $\pi(t,j) \in \{1,\dots,360\}$ with the Levenberg-Marquardt backpropagation and a total number of $N^{\text{ANN}}_{\text{anneal}} = 10$ annealing steps. As described in Section 4.1, the different network candidates are validated by considering the corresponding residuals. These deviations are particularly small for $N^{\text{ANN}}_{\text{neurons}} = 10$. Thus, the associated network $\mathcal{N}(10)$ is used to compute an estimate of $\pi^{\text{syn}}(t,j)$.

It is apparent that the model is able to accurately estimate the artificial data associated with the considered dynamical system although essential information on the relevant driving forces was not available for the parametrization procedure.

3 Note that the stationary, homogenous case is considered as well.

Due to the high number of regimes required to qualitatively describe the artificial system, any prediction of $\pi^{\text{syn}}(t,j)$, i.e., $t \in \{361,\dots,400\}$, is particularly complex. Consequently, prediction steps of depth 1 are considered. Essentially that means that each prediction $\hat{\Gamma}(t,j)$ of the regime probabilities for $t > 360$ is updated for the next prediction step. Under the assumption that new data $\pi(t+1,j)$ can be retrieved, an update, based on the maximum-likelihood principle, is used (see (3.43)):

$$\gamma_k^*(N_T+1,j) \tag{4.26}$$
$$= \begin{cases} 1 & \text{if } k = \underset{h}{\operatorname{argmin}}\, g(\pi(t+1,j),\dots,\pi(t-N_M,j),\theta_h(u(t,j))), \\ 0 & \text{otherwise.} \end{cases}$$

Note that $\gamma_k^*(N_T+1,j)$ is assumed to be optimal, hence the superscript asterisk. An alternative update strategy, motivated by Bayes' theorem, conditioned on the retrieval of new observations, has recently been proposed and validated in [31]. The resulting out-of-sample approximations (i.e., one-step predictions), computed with the model $\mathcal{M}^{\text{Markov}}(4,15,0,u^{\text{syn}}(t,j))$, are promising and suggest that it is possible to compensate lack of information via the proposed non-stationary, non-homogenous parameters.

Further, the two plots of Figure 11 reveal that the feasibility of the considered network $\mathcal{N}(10)$ highly depends on the location. This phenomenon can be explained with the fact that the implicit external factors $u_e^{\text{unres}}(t,j)$ given in (4.22) are dependent on location, and thus the corresponding effect differs for different cells j. Essentially that means that the quality of the estimates of the state probabilities via ANN is considerably reduced without the additional information on the influencing quantities (see upper panel of Figure 11). The reason for that is that model classes such as ANN as well as SVM have time-independent parameters (i.e., the weights and the bias of a neuron are global parameters) and are thus intrinsically stationary. Yet, it is possible to calculate good approximations with the considered network (see lower panel of Figure 11) as long as the dynamics are not effected as much by the unresolved factors.

As the influence of implicit external factors is taken into account, the approximations computed by means of the Markov model have a high accuracy (independent of the location) for the considered data set. Concluding, the introduced framework can compete with standard data-analysis techniques. Moreover, for ill-conditioned problems that require to capture the effects of unobserved external factors, the non-stationary, non-homogenous structure of the Markov model is a more reliable option.

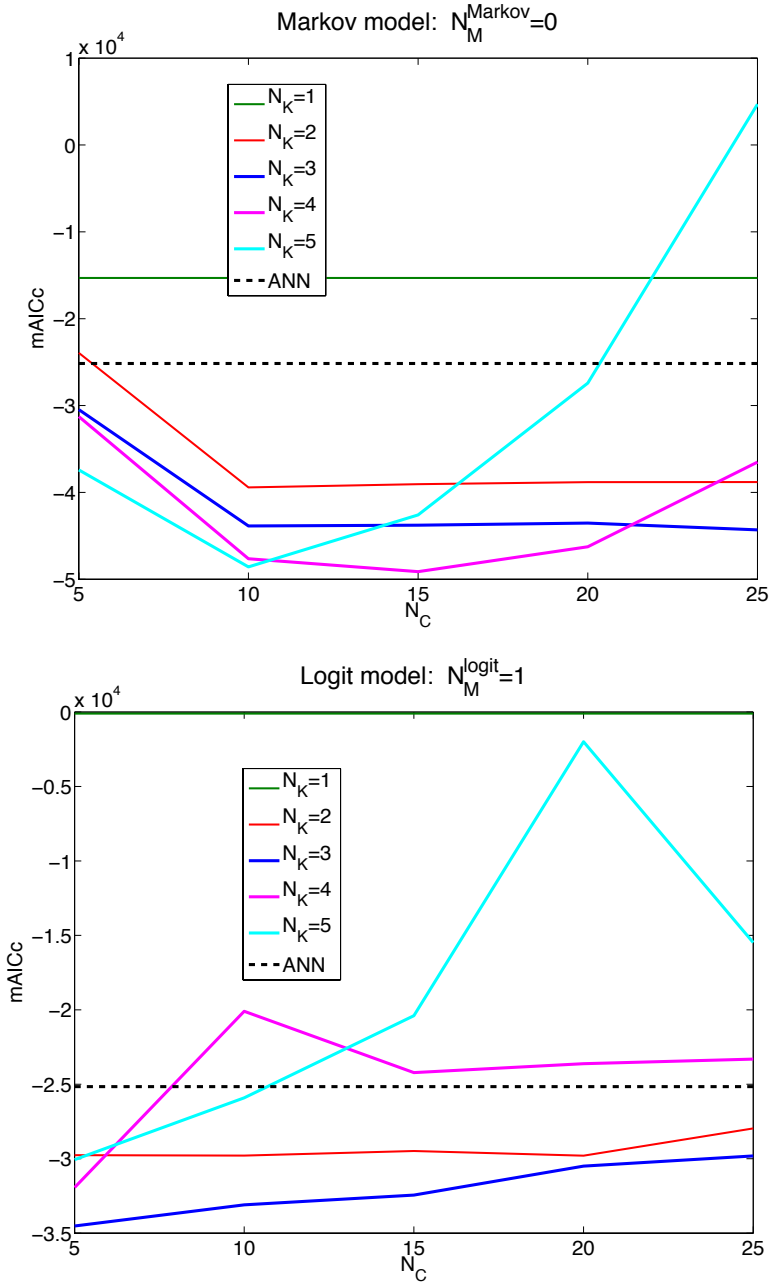

Figure 10: *The mAICc values for different models $\mathcal{M}^f(N_K, N_C, N_M, u^{syn}(t,j))$ for $N_K \in \{1,2,3,4,5\}$ and $N_C \in \{5,10,15,20,25\}$ are visualized. More precisely, the results for $N_M = 0$ and $f = Markov$ can be seen in the lower panel, whereas the values corresponding to $N_M = 1$ with $f = logit$ are shown in the upper panel. Additionally, the mAICc value for the optimal (with respect to residuals) network $\mathcal{N}(10)$ is displayed.*

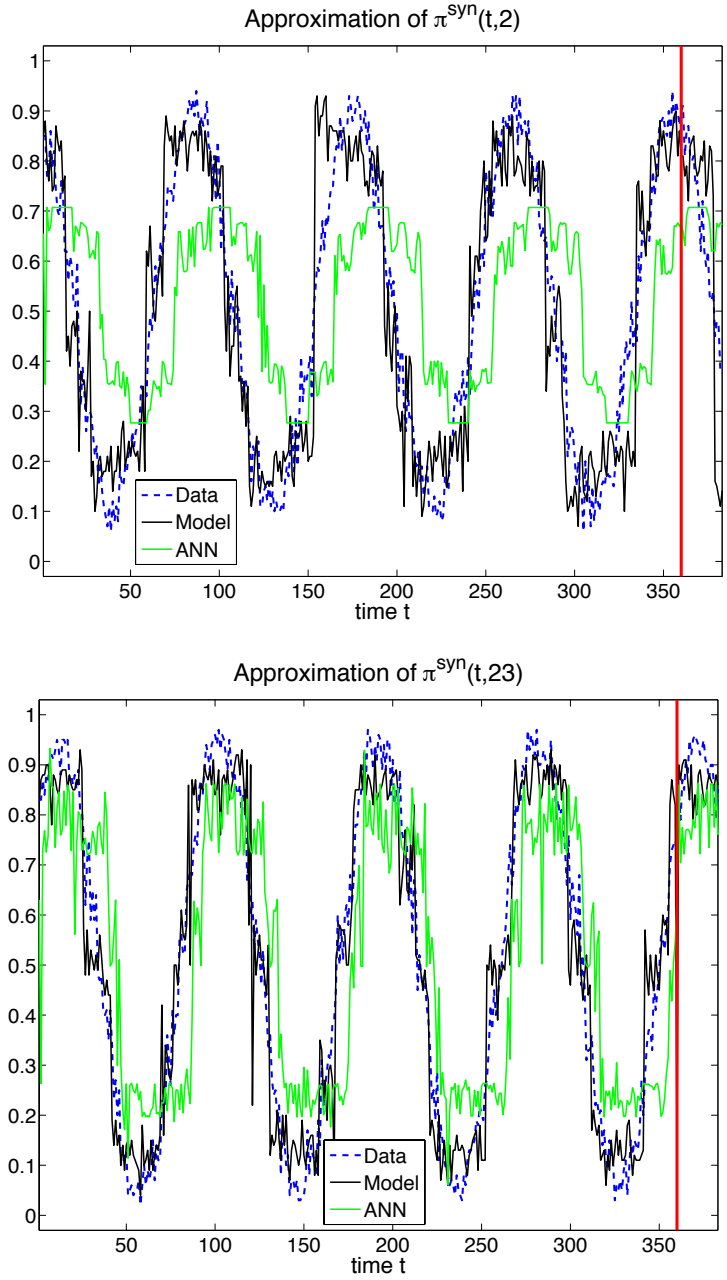

Figure 11: *Two approximations, computed via different modeling approaches, of the synthetic data $\pi_1^{syn}(t,j)$ for two exemplary locations, i.e., $j = 1$ (see upper panel) and $j = 23$ (see lower panel), are visualized in this figure. The data approximation, determined by means of the memory-less model $\mathcal{M}^{Markov}(4, 15, 0, u^{syn}(t,j))$, is shown in form of a black line, and the estimation in green is generated according to the network $\mathcal{N}(10)$. Moreover, the corresponding artificial time series $\pi_1^{syn}(t,j)$ is displayed as a dashed blue line. Note that the begin of the out-of-sample approximations is marked with a red vertical line at $t = N_{T_{train}} = 360$.*

5

ARCTIC SEA ICE APPLICATION

In this chapter the proposed methodology is applied to arctic sea ice observations in order to infer an appropriate data-based model describing the arctic ice dynamics manifesting in the sea ice extent. This particular application was chosen due to the fact that, firstly, it is a multidimensional data set in both components, i.e., in space as well as in time, suiting the theoretical discrete setting of the model.

Secondly, the underlying physical dynamics and interactions, causing the de- or increase of sea ice, have a complex nature and are not usually available in form of measurements. Thus, the functionality of the parametrization technique can be specifically tested on a complex system with missing information. Subsequently, the theoretically verified strength of the method can be practically investigated.

Thirdly, the vast arctic sea ice loss in recent years poses a major thread to the current global climate, and thus it is particularly important to understand the associated dynamics.

Before the analysis set-up is described, the arctic data is discussed. In that context, a framework employed to project the data to a hexagonal lattice is outlined. After giving information on the computational details, the optimal model parameters (selection via the proposed information criterion) are interpreted. Further, the out-of-sample performance of the corresponding model and the statistical impact of the explicit external factors are considered.

5.1 DATA

The spatial extent of the arctic sea ice coverage can be measured via satellite. The resulting data product usually consists of the sea-ice concentration values, i.e., the percentage of sea ice on the ocean surface in each regarded grid-box $j \in \{1, \dots, N_J\}$. The data can be assumed to be the distribution of state probabilities $\pi(t, j)$ of a microscopic process $\sigma(t, j, l)$ describing the

aggregate state of microscopic locations, i.e., switching between two states $s_1 =$ solid and $s_2 =$ liquid.

Data availability with respect to the size of the grid cells, the length of the observed time period, the time-wise frequency of measurements, and the covered areas of the arctic region varies a lot. In particular, the corresponding explicit external factor observations might not coincide with respect to time or space with the considered measurements of the state probabilities $\pi(t, j)$. Thus, some compromises are made to be able to work with a consistent data set.

The observations considered in this manuscript were made publicly available by the *National Snow & Ice Data Center* [20][1]. The regarded data covers a period of 16 years (from January 1989 to December 2004) in biweekly time steps (i.e., $N_T = 384$ time steps in total). It spans over an area starting with latitude values 45° N (southernmost) and going up to 90° N (northernmost) covering the entire circle, i.e., longitude lines from westernmost 180° W to easternmost 180° E.

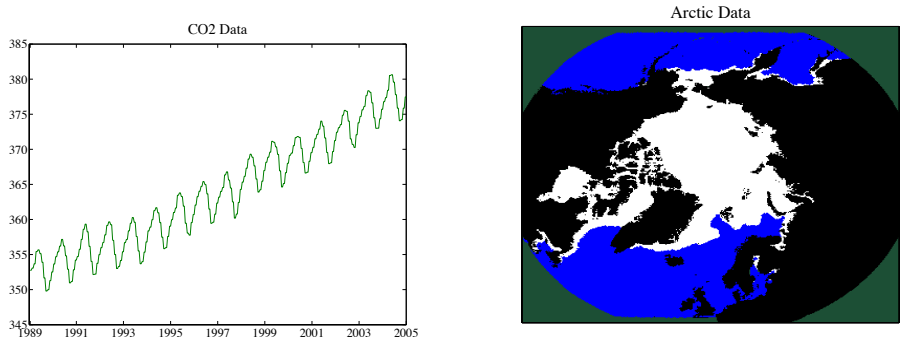

Figure 12: *The panel on the left displays the CO_2 data values in ppmv, considered for the computations with the non-stationary, non-homogenous Markov regression. The collection is derived from in situ air measurements at Mauna Loa, Hawaii, USA. The image on the right shows a rounded excerpt of the satellite data in EASE-Grid format. More precisely, a visualization of the data, observed in the first two weeks of January 1989, is displayed. The sea ice concentration values above and equal to 50 are displayed in white, values below 50 are shown in blue, and the land cells are colored in black. Note that the additional green cells correspond to areas not covered with the satellite.*

The information can be downloaded in an EASE-Grid North Azimuthal format[2] [16] with 361×361 grid cells and a resolution of 25 km, i.e., each

1 The arctic sea ice coverage data can be downloaded on `http://nsidc.org/data/docs/noaa/g02172_nic_charts_climo_grid/index.html`.

2 A detailed documentation on the family of NSIDC EASE-Grid formats can be found on `http://nsidc.org/data/ease/ease_grid.html`

cell covers an area of 25 square km[3]. The information associated with each sea cell, i.e, a cell covering an area of the arctic ocean, is the percentage value of surface sea ice observed on the corresponding 25 square km. The cells that cover areas on continental land masses are specifically labelled and are not affiliated with any additional ice information.

An image of the 361×361 squared lattice (EASE-Grid North Azimuthal format) displaying the given data info can be seen in the panel on the right of Figure 12. It shows the arctic ocean with the existing ice cover in the first weeks of January 1989, part of the surrounding continents (mainly Canada, the United States and Russia), prominent Bays (e.g., Hudson Bay), and parts of the adjacent oceans. Additional to the ensemble observation $\pi(t, j)$, a set $\mathcal{E}$ of measurable external forces is considered for the parametrization of microscopic sea ice dynamics described by the process $\sigma(t, j, l)$. In the following subsection, information on the type and the sources of this set of explicit external factors is given.

5.1.1 *External factors*

In some cases it might not be a problem to get access to qualitative ensemble data, i.e., to obtain $\pi(t, j)$, but only very little measurements exist of influencing factors $\bar{u}(t, j)$. To some degree, lack of information is already taken into account via the non-stationary, non-homogenous model parameter $\Gamma(t, j)$, i.e., implicit external quantities are reflected in the explicit dependency on location j and time t. Nevertheless, in order to obtain a qualitative characterization of the underlying process $\sigma(t, j, l)$, it is necessary to collect as much information as possible in form of observations about possible influencing quantities. Thus, in this thesis the considered time span is specifically chosen according to the availability of certain explicit external factors.

Although many of the explicit external factors are available for many time steps, there is usually a limitation concerning the area of the arctic covered, e.g., temperature measurements etc. are usually available close to the coast but are more scarce closer to the pole. Oceanographers also name salinity, wind, ocean currents, and temperature below the ice among the relevant factors [107]. Yet, there are no consistent observations of these quantities available for the considered arctic area, matching the time component of the regarded coverage data.

3 Though it is important to stress that the number of locations considered for the computations with the regression framework is much smaller, i.e., 198 locations in total.

As already mentioned above, the collection of available explicit external factors is denoted $\mathcal{E}$. The factors in $\mathcal{E}$ associated with the considered ice dynamics are now described.

Firstly, as discussed in Section 3.3, the spatial neighborhood of a cell plays an important role. Existing correlations have to be taken into account in the model. Consequently, the states of adjacent cells are included in form of an exterior force. As a hexagonal lattice is considered, each cell has six neighboring cells all sharing an edge with the respective cell.

A distinction is made between the influence of adjacent ice concentration and surrounding land masses. The major importance of land cells has recently been discussed in [35]. It is suggested that differences in the arctic and the antarctic ice growth can be affiliated with the surrounding land masses. Thus, although land cells are not included in the total number of N_J locations, they should be involved in the parametrization process. The effect of landmasses on the ice shelfs close to the coast is studied by adding mean surrounding land percentage values as an explicit external factor. In detail, neighboring sea ice concentration and adjacent land masses are considered as a mean of the data of the six neighbor cells, i.e.,

$$\text{neigh}_{\text{ice}}(t,j) := \underset{\text{r}\in\text{ neigh(j)}}{\textbf{average}}(\pi_1(t-1,r)) \tag{5.1}$$

and

$$\text{neigh}_{\text{land}}(t,j) := \frac{1}{6}\sum_{\text{r}\in\text{ neigh(j)}} \textbf{land}(r) \quad \forall t, \tag{5.2}$$

where $\textbf{land}(r)$ equals 1 if r is a land cell and 0 otherwise.

Obviously, the temperature plays an important role in the dynamics of sea ice. Hence, measurements of temperature values in each location j and for every time step t, denoted $\text{temp}(t,j)$, are considered as an exterior factor, i.e., $\text{temp}(t,j) \in \mathcal{E}$. A representative of the data for fixed t (measured in the summer of 1989) is shown in the graphic on the left of Figure 19. The continental land masses are added to the graphic in form of black cells in order to give a better understanding of the considered area.

Note that temperature values assigned to the geographical coordinates are already projected via a geodesic DGGS onto a hexagonal raster. The details of this particular transformation and the corresponding software are discussed in the following section.

Further, one distinguishes between local and global external influences. Put differently, observable quantities $u_e(t,j)$ can explicitly depend on the location, i.e., be local, or take the same value for all cells, i.e., be global.

For instance, the fourth quantity contemplated for the computations is the atmospheric CO_2 concentration in the air, measured on a single location that is not in the arctic region.

The $CO_2(t, j)$ observational values are the same for all locations $j \in \{1, \ldots, N_J\}$, and thus CO_2 is a global influencing factor. The particular data set used (see panel on the left in Figure 12) is a collection[4] of in situ air samples that have been measured at Mauna Loa, Hawaii, USA and are given in parts per million by volume (ppmv)[5].

Note that it is also possible for the external factors to be global in a time-wise sense. For instance, the mean of the surrounding land masses $\text{neigh}_{\text{land}}(t, j)$ does not change in time but changes for each location j.

Recent research has revealed that the arctic sea ice concentration is influenced by global teleconnections such as northern atlantic oscillation (NAO) as well as the arctic oscillation (AO) index [4, 91, 57]. Thus, these two climate phenomena are also considered as resolved external factors, i.e., NAO, AO $\in \mathcal{E}$.

The NAO data set[6] considered in this manuscript is computed with a procedure using rotated principal component analysis (RPCA), described in [6] (see upper panel of Figure 13). The loading pattern of the NAO is defined as the first leading mode of rotated empirical orthogonal function (REOF) analysis of monthly mean 500 mb height during the 1950-2000 period.

The used AO data[7] is a monthly mean of the daily AO index, which is a projection of 1000 mb height anomalies occurring north of 20° N latitude onto the considered loading pattern of the AO[8] (see lower panel of Figure 13). Summarizing, the set of potentially relevant explicit external factors is set to be:

$$\mathcal{E} := \{\text{neigh}_{\text{ice}}, \text{neigh}_{\text{land}}, \text{temp}, CO_2, \text{NAO}, \text{AO}\}. \tag{5.3}$$

4 The entire CO_2 data set is available on `http://cdiac.ornl.gov/ftp/trends/co2/maunaloa.co2`.

5 It is important to stress that ppmv is not officially a unit, but commonly used for greenhouse gas measurements. It refers to the millionth part (ppm), i.e., 10^{-6}, such as percentage % values refer to the hundredth part, i.e., 10^{-2}. One further distinguishes between mass fraction and mole fraction (also described as "by volume"), which means the ratio of molecules of the regarded substance with respect to all the molecules in the considered volume. Since (3.35) and (4.5) are affine-invariant linear transformations, without any loss of generality, all of the external factors $u_e(t, j)$ can be made dimensionless (or unit-less) and can all be transformed to some uniform interval (e.g., $[-1, 1]$).

6 The NAO data set is available on `http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/norm.nao.monthly.b5001.current.ascii.table`.

7 The AO data set considered in this manuscript is available on `http://www.cpc.noaa.gov/products/precip/CWlink/daily_ao_index/monthly.ao.index.b50.current.ascii.table`.

8 The loading pattern is defined as the leading mode of Empirical Orthogonal Function (EOF) analysis of monthly mean 1000 mb height in the time interval 1979 to 2000.

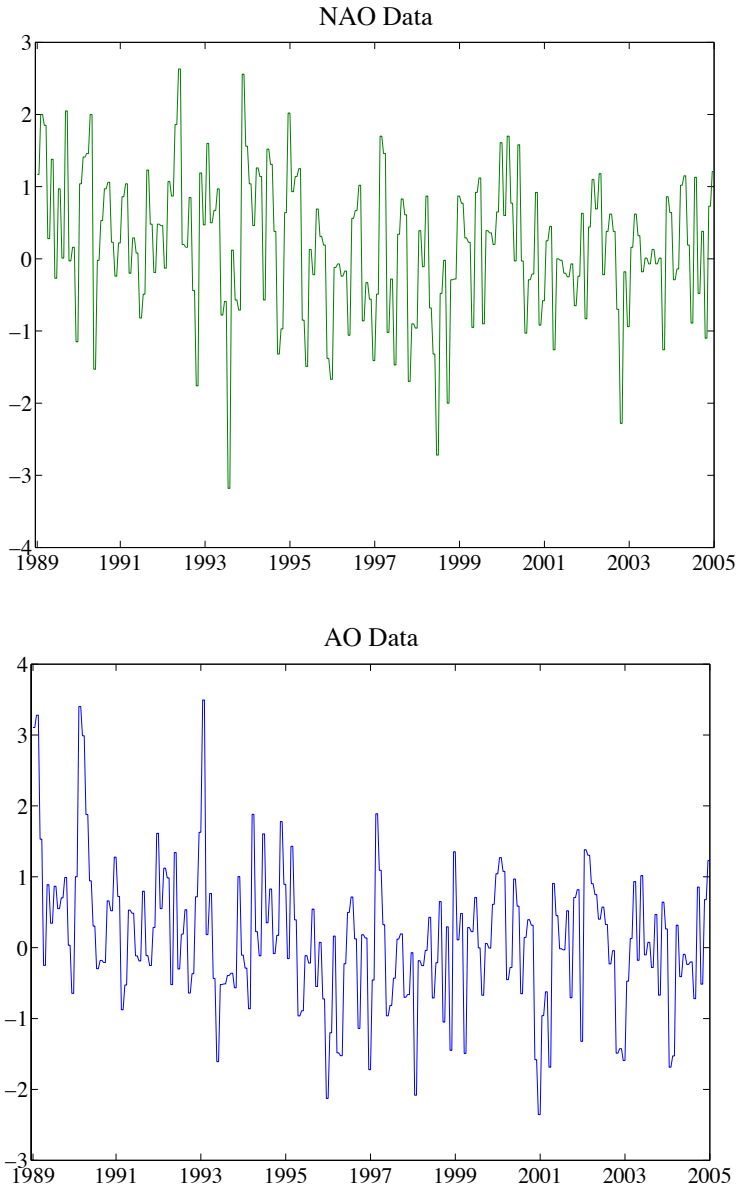

Figure 13: *Considered data set of NAO and AO index in the time interval January 1989 to December 2004.*

As mentioned above, a hexagonal lattice is considered for the following computations. The details of the projection of the data onto a hexagonal lattice are discussed in the following section.

5.2 DISCRETE GLOBAL GRIDS

Modeling or monitoring of global dynamical processes demands partitioning of the surface of our planet into discrete grid cells. A discrete partition of the entire Earth's surface is referred to as *Discrete Global Grid System* (DGGS). Ideally a DGGS has certain useful properties helping to improve analysis and measuring techniques. A suggestion of such properties is listed in the Goodchild Criteria [64], named after Michael Goodchild [43], who formulated an earlier version.

The top five of the fourteen items in the list can be summarized as follows: The chosen discrete grid allows a complete tiling of the globe without overlaps. Further, the grid cells should have equal surface area, be of the same shape, and topology (i.e., have the same number of edges and vertices). Moreover, it is required that the processes comprised by a grid cell have similar nature, which is referred to as *compactness*.

Since no existing global grid fully meets these criteria, one usually settles for the best option for the currently considered environmental phenomenon. In fact, many standardly employed partitions of the Earth's surface, induced by the geographical coordinate system given by latitude and longitude values, do not have equal-area cell regions. This problem becomes even more pronounced, i.e., area and shape of the cell are more distorted, when approaching the North Pole or conversely the South Pole from the equator. Furthermore, many data sets, especially the ones collected priorly to the computer era, are associated with latitude and longitude values, thus, DGGSs based on geographical coordinates and the corresponding processing algorithm are well established and popular in a wide range of applicational areas.

Although in general it is not possible to design a DGGS that is optimal for all applications, one option is to improve the construction of such a system in order to meet more or specific criteria on Goodchild's list for an ideal DGGS. More specifically, a new class of geospatial data structures, referred to as geodesic DGGSs, has been proposed as an alternative to existing approaches.

As already discussed in Section 3.4, a honeycomb lattice is considered in this thesis. Thus, the aim is to project the regarded observation data to a hexagonal grid. Unfortunately, a tiling of the surface of the Earth with solely hexagonal cells is not possible [98]. Nevertheless, a raster with mainly honeycomb cells, associated with points on Earth, can be designed via the mentioned geodesic DGGSs, which are introduced and surveyed in the following subsection.

5.2.1 *Geodesic discrete global grid systems*

As an alternative to the commonly used planar raster of quadrangles, induced by the latitude-longitude graticule, promising results have been proposed on the basis of regular polyhedra. The conceptual idea is to exploit the topological equivalence of the regular polyhedra and the 2-sphere (which can later be related to the Earth's surface). This matter of fact can be explained best via Euler's polyhedron theorem which states that all convex polyhedra fulfill the following formula

$$\#\text{vertices} + \#\text{faces} - \#\text{edges} = 2. \tag{5.4}$$

It is possible to derive from Equation (5.4) that there are exactly 5 *platonic solids*, i.e., regular, convex polyhedra [106]. The regular polyhedron with triangular faces, called the *icosahedron*, for instance, is one of the platonic solids and has 20 faces. Note that there is no regular, convex polyhedron with hexagonal faces.

Considering Euler's polyhedron formula in a more general context, it states that the surface of a convex polyhedron has *Euler characteristic* equal to 2. Subsequently, the surface of a convex polyhedron is homeomorphic to the surface of the 2-sphere. Due to the fact that a spherical or ellipsoidal surface is generally regarded to be a good surrogate for the Earth's surface, this ansatz allows to construct a variety of different DGGSs, which, following [98, 120], are referred to as *geodesic DGGS*[9].

A list of five characteristic design choices specifying a considered geodesic DGGD is given in [98]. The first choice involves to pick a regular base polyhedron. The authors of [98] focus on the five platonic solids. Nevertheless, it is also possible to consider other convex polyhedra as they also fulfill Euler's polyhedron formula and are thus also homeomorphic to the surface of the 2-sphere. For instance, in order to construct a partition of the Earth's surface, predominately consisting of hexagonal cells, an alternative option is to regard the convex truncated icosahedron, which is a polyhedron with two different types of regular polygons as faces.

The truncated icosahedron has 12 regular pentagonal and 20 regular hexagonal faces and has a surface homeomorphic to the 2-sphere surface. The name already hints at the derivation from the icosahedron. The construction of this particular truncated polyhedron is achieved via cutting off (truncating)

9 The name stems from the fact that many of the systems based on regular polyhedra have been inspired in some way by the scientific research of Buckminster Fuller, who designed the geodesic dome.

12 specific vertices at one third of the corresponding edges, resulting in 12 pentagonal faces, replacing the former vertices and forming 20 regular hexagonal instead of the triangular faces. Then it is possible to regard the spherical truncated icosahedron, which is a partition of the surface of the sphere into spherical polygons using great arcs. This particular spherical truncated polyhedron is used for most of the modern soccer ball designs, hence the spherical truncated icosahedron is a commonly known tessellation of the sphere.

However, since an equivalent partitioning can be constructed using a regular icosahedron as the base polyhedron of the considered geodesic DGGS, the focus in this thesis, following the proposed structure in [98], is exclusively on the platonic solids. More specifically, the icosahedron is considered. This design choice is sensible with respect to a hexagonal tiling and has a relative small distortion concerning the spherical transformation of the polyhedron compared to the other platonic solids with bigger faces, e.g., the tetrahedron or the cube [120].

The unfolded planar icosahedron is commonly considered in a partition of 10 quadrilaterals, each formed by a pair of triangular faces. In the remainder of the thesis, the quadrilaterals are indexed as displayed in Figure 14, where the surroundings of the respective quadrilateral on an unfolded icosahedron are marked by bold black lines and the former triangles are indicated via thin dashed lines, also in black.

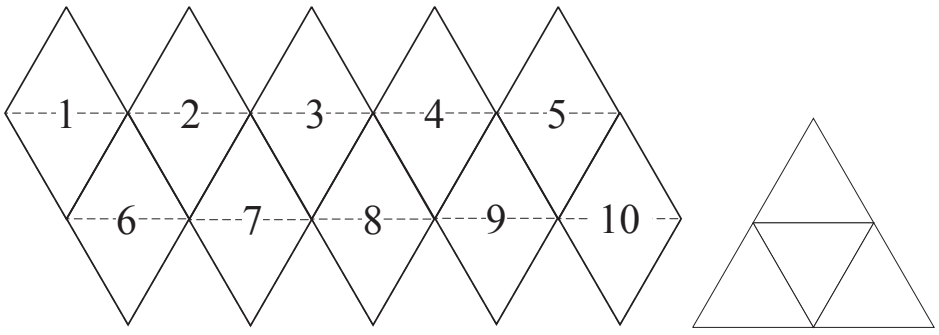

Figure 14: *The graphic on the left shows the planar unfolded Icosahedron, where pairs of triangles are combined to form ten quadrilaterals. The specific numbering used in this figure is deployed for the entire projection procedure and the corresponding discussion. On the right an aperture 4 partition of a triangle is displayed.*

Secondly, a lattice with mostly honeycomb shaped cells is regarded. In order to obtain a dominantly hexagonal raster, the 20 triangular faces of the icosahedron (see Figure 14) are subdivided into hexagons by cutting off each vertex at one third of the edges, leaving small rest areas, which

together form 12 pentagonal cells. This procedure is demonstrated by means of one triangular face in Figure 15 (see first triangle from the left). The areas forming the pentagonal cells are colored in grey.

Thirdly, it is necessary to decide upon the orientation of the icosahedron relative to the Earth. The orientation is defined by fixing the geographical coordinates of a vertex of the respective platonic solid and additionally set the azimuth information of an adjacent vertex. The benefits of a choice are substantially dependent on the application. For instance, the polyhedron can be oriented so that a certain continent or country is well placed (e.g., on one of the faces) [119].

Another approach is to align the edges, vertices, and faces with important quantities such as the prime meridian, the poles, or the equator [96, 112, 119, 121]. For example, a popular orientation for a icosahedral base polyhedron is to assign a vertex to each of the poles and to place one of the edges (connected to the vertex at the North Pole) of the platonic solid so that it is aligned with the prime meridian [96, 112]. The disadvantage, however, is that for this common orientation the icosahedron is not symmetrical about the equator, which is not a desirable property for some applications [50, 51]. In order to change the orientation to be symmetrical about the equator, the icosahedron is rotated by 36°.

The considered parametrization of processes in the arctic circle area is not necessarily affected by the orientation. Yet, it is easier to approach the corresponding implementation and the visualization of the respective results for an orientation that places the entire arctic area on a small number of quadrilaterals. Thus, the orientation defined by one vertex at 11.250E longitude, 58.282525590N latitude and an adjacent vertex at an azimuth of 0.00 is considered due to the fact that most of the arctic circle area is projected on only two of the quadrilaterals assembling the spherical icosahedron.

Alternatively, the Dymaxion orientation of R. Buckminster Fuller, i.e., one icosahedral vertex at 5.245390W longitude, 2.3008820N latitude with an adjacent vertex at an azimuth of 7.466580, leads to a similar convenient placement of the arctic region.

The fourth design choice involves an appropriate hierarchical spatial partitioning technique that allows to create different grid sizes of the geodesic DGGS. In that context, the *aperture* of a DGGS is the factor describing the area ratio between the current and the next finer resolution [98]. For example, regarding an equilateral triangular face, it is possible to divide each triangle into four equilateral triangles, thus the cell area is reduced by a factor four (see Figure 14), i.e., defining an aperture 4 triangle hierarchy. In cases where

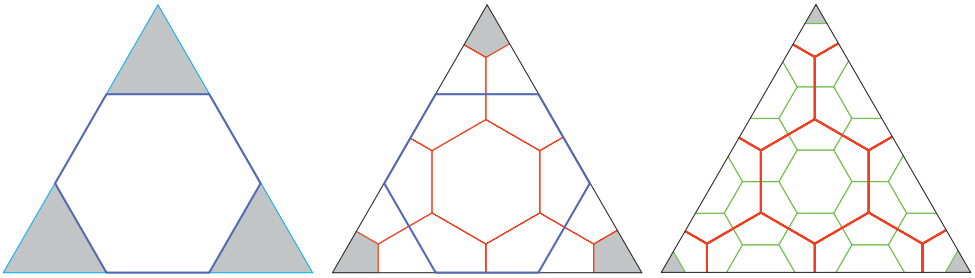

Figure 15: *The figure displays three triangles with a partitioning predominately consisting of hexagonal cells. The hexagons are divided in the aperture 3 sense. The first grid resolution is colored in blue and originates from the partitioning of a base triangular face. The grey areas correspond to parts of the 12 pentagonal cells existing in each grid resolution. The second resolution is displayed in the center and in the third triangle from the left and is colored red. The finest grid shown is associated with resolution 3 and is colored green (see third triangle from the left).*

the underlying polyhedron is divided to have more than one type of polygon as its faces, the aperture is defined for the dominating cell shape.

As it is not possible to entirely tile a hexagonal cell with smaller hexagons, any partitioning used to define a hierarchy of finer grid resolution is more complex and does not allow a straightforward characterization of relations between cells in different grid sizes. This leads to computational difficulties concerning sensible hierarchical location coding and run time as well as memory efficiency. These issues are discussed and addressed in detail in [97].

In this thesis, aperture 3 hexagonal cells are considered, meaning that the area of a hexagonal face in the next finer resolution has a third of the area of the current coarser grid. This particular partitioning of the hexagonal cells to a smaller honeycomb structured raster is shown in Figure 15. The first triangle from the left in Figure 15 displays a resolution 1 hexagon in blue which is divided into a resolution 2 grid (red lines), shown in the center triangle. Finally, a resolution 3 hexagonal raster (green lines) is displayed in the third triangle.

The possible sizes of aperture 3 hexagonal cells in the employed DGGRID version 3.1b ranges from resolution 1 relating to 20 hexagonal cells and 12 regular pentagonal cells to resolution 18 with 3,874,204,880 hexagons and 12 pentagons. Note that the number of pentagons for each resolution is fixed to be 12. For the computations in this thesis, resolution 6 is considered. In detail that means that each of the 10 quadrilaterals has 27 times 27 cells including one pentagonal cell and two additional connecting pentagonal

cells. The total number of cells in resolution 6, tiling the entire surface of the Earth, is 7,292, where each cell has a size of approximately 69,968 square kilometers.

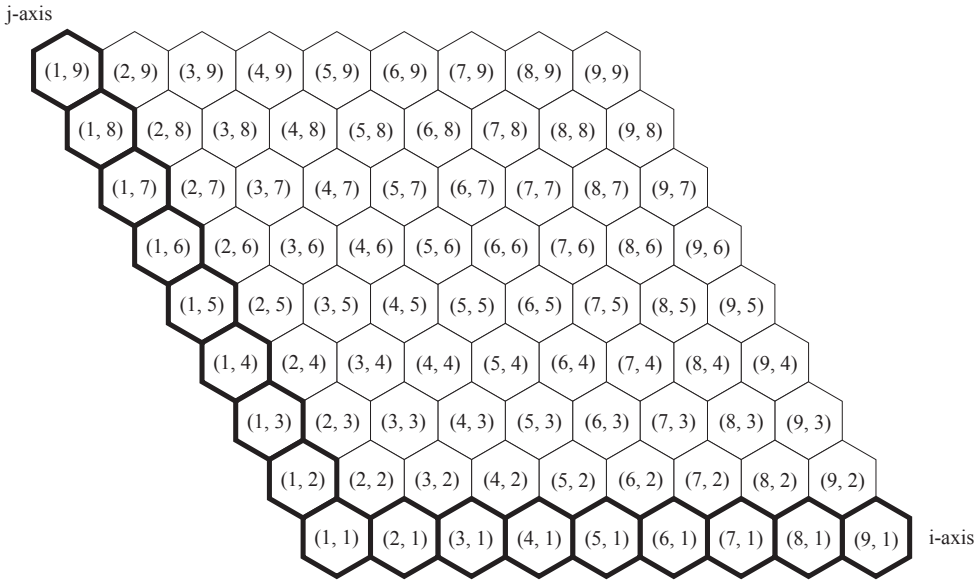

Figure 16: *The figure demonstrates the employed address assignment for each location for a specific quadrilateral q by means of example in resolution 4, i.e., the size of every quadrilateral is* 9×9*. Note that the pentagonal cell (at origin* $(1,1)$*) is visualized, for structural reasons, as a hexagon as well.*

Further, a transformation method projecting the planar base polyhedron onto the corresponding spherical/ellipsoidal surface needs to be specified. The focus is on the subdivision of the polygonal faces as the corresponding spherical counterpart is desired to be similar with respect to the partition. One distinguishes between two standard approaches [64]. One ansatz is to employ a projection that directly relates a subdivision of the spherical/ellipsoidal surface to the considered partition on the planer polyhedral faces. Alternatively, it is possible to project a chosen partition on the planar faces of the polyhedron via an inverse map transformation to the sphere or the ellipsoid, i.e., to an Earth surrogate.

As discussed in [98], any projection with the property to map straight-line planar face edges to the great-circle arc edges of the corresponding spherical face can be used. One example function fulfilling this requirement is, for instance, the Fuller Dymaxion projection [38, 44]. The quality or properties of the different transformations usually vary in terms of area and shape distortion. For the purposes of this thesis, it is sensible to consider an equal-area projection. Thus, the Icosahedral Snyder Equal Area (ISEA)

projection [105] is used as an inverse transformation between an appropriate partitioning of the Earth surrogate and a planar icosahedron.

Summarizing, these five specific design choices result in a grid, called the Icosahedral Snyder Equal Area aperture 3 Hexagon geodesic DGGS (ISEA3H geodesic DGGS). A free software package, made available by Kevin Sahr[10], one of the developers of a particular set of geodesic DGGS, was employed in pursuance of projecting the considered arctic sea ice data associated with latitude and longitude coordinates onto a lattice in the plane with dominantly hexagonal cells.

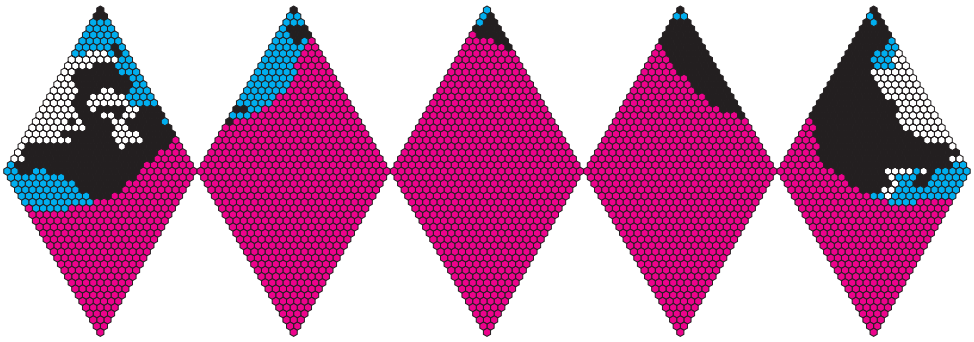

Figure 17: *The image shows a projection of the rounded data observed in January 1989 onto a planar icosahedron with a hexagonal tiling in resolution 6. The cells colored in pink correspond to unknown data points and the black cells are associated with land. Further, the state assignment of ice (white) or of water (blue) is determined by means of the rounded observations.*

The software allows different address assignment formats for the output, and the reader is referred to [97] for a detailed description of location coding and a discussion on the addressing of locations. For example, it is possible to get information about the index of the quadrilaterals with corresponding (i, j) coordinates (see *output_address_type* Q2DI and exemplary address assignment in Figure 16)). Then the hexagonal cells of one quadrilateral are indexed as shown in Figure 16. For resolution 6, the indices i and j take values from 1 to 27.

As the locations of the regarded data are geographically placed on the Northern Hemisphere, not all of the quadrilaterals contain data points. In fact, the main area considered is projected onto only two quadrilaterals. A snapshot of the data corresponding to the first two weeks of January 1989, projected onto a planar icosahedron with a hexagonal tiling via the specified geodesic DGGS, is displayed in Figure 17. Most of the relevant

10 The software, a corresponding handbook, and most of the relevant publications can be downloaded from `http://discreteglobalgrids.sqsp.com/software/`.

data points are projected to the first and the fifth quadrilateral. This can be explained with the fact that two of the 12 pentagonal cells are not contained in the 10 quadrilaterals. Note that the indices of the quadrilaterals in the software package start with zero and go up to 11. The sea ice coverage information assigned to the pentagonal cell (i.e., the cell corresponding to the Northern Hemisphere) is not displayed in Figure 17. The cells colored in pink correspond to unknown data points, i.e., there is no observation data available due to the fact that the area on Earth is too far away from the arctic circle and hence irrelevant.

In order to work with a sensible number of locations, only cells of quadrilaterals one and five are considered in the following. In fact, 198 cells of the 1458 cells of the two quadrilaterals are regarded. This number is a result of discarding all the cells without information, i.e. pink ones, all the land cells, i.e., grey ones, and finally some cells that are assigned a probability to be in state water for the entire season. To be precise, a small number of cells with a potential ice state were also not considered due to the fact that the aim is to regard a lattice of connected grid cells.

As the regarded lattice is finite, it is important to decide what kind of boundary conditions should be considered. A common approach is to work with periodic boundaries. Since the number of cells covered by the satellite is greater than the number of cells considered for the computations, it is possible to use the existing information of adjacent locations that are not associated with one of the N_J cells. The key advantage is that neighbor states do not have to be artificially created but are already available.

5.3 PARAMETER IDENTIFICATION AND RESULTS

In the following, the numerical settings for the parametrization of the considered arctic sea ice dynamics with the introduced non-stationary, non-homogenous Markov regression are discussed. Then the selection of optimal parameters is outlined and the resulting model is interpreted in terms of the application. This also entails to determine the statistical impacts corresponding to the considered explicit external factors. Afterwards, the performance of the inferred model is validated by comparing its results with the actual data.

5.3.1 *Analysis set-up*

As the considered explicit external factors $u_e(t,j)$ have different scales, it is prudent to unify the entries of $u(t,j)$, i.e., to rescale them so that

$$-1 \leq u_e(t,j) \leq 1 \quad \text{for all } t, j \text{ and } e. \tag{5.5}$$

Consequently, the unit of a respective factor is irrelevant for its relative influence magnitude. Approximately 94% of the data is used for the purpose of determining a set of model parameters describing the underlying arctic dynamics. In detail, this training period spans from January 1989 to December 2003 (i.e., $t \in \{1, \ldots, 360\}$). The remaining portion ($\approx 6\%$) of the data is used to validate the predictive skills of the trained models. Summarizing, the arctic sea ice coverage observations of 2004 are compared to out-of-sample approximations computed with the obtained model (i.e., $N_{T_{\text{train}}} = 360$).

In order to determine the optimal (with respect to the considered information criterion) model parameters on the basis of the data, the Markov regression runs are executed for different values of local regimes N_K and the maximal number of possible transitions N_C of the affiliation process, i.e., $N_K \in \{1,2,3,4,5,6,7\}$ and $N_C \in \{5,10,20,30,40,50,60,70,80\}$. Further, the model parameters are computed for all subsets of the considered set $\mathcal{E}$, i.e., 63 runs are completed to determine the best fit for all possible combinations of the six resolved quantities. Moreover, the memory-less special case, i.e., $N_M = 0$, where the state probabilities of the next time step are independent of the current state probabilities (see (3.42)), is considered additionally to the standard Markov model (i.e., $N_M = 1$). This results in a total of 7938 Markov models $\mathcal{M}^{\text{Markov}}(N_K, N_C, N_M, u(t,j))$, which have to be parametrized and tested for the given data.

Further, the dynamics underlying the affiliation process $\Gamma^*(t,j)$, associated with the model that attains the lowest mAICc value, have to be determined in pursuance of considering the out-of-sample performance of the model. Following the idea of a self-containing predictive model, Algorithm 4 is employed to compute different stationary and homogenous models

$$\mathcal{M}^{f^\Gamma,\Gamma}(N_K^\Gamma, N_C^\Gamma, N_M^\Gamma, u^\Gamma(t,j)). \tag{5.6}$$

Note that $f^\Gamma \in \{\text{Markov}, \text{logit}\}$ and that the entries $u_e^\Gamma(t,j)$ of the vector of explicit external factors are elements of the set $\mathcal{E}^\Gamma$. The optimal model describing the obtained process Γ^* is again selected via the mAICc.

As already mentioned in Section 3.6, an alternative ansatz to choose an appropriate candidate in the set of different models is to apply cross validation, i.e., to compare the respective out-of-sample performance of the different models. However, for the given model structure this entails to determine at least 63504 different models (due to the many possible combinations for the characterization of Γ^*), which then have to be compared in terms of approximation quality. Summarizing, the cross validation approach for this example is not computational sensible. Yet, in general, cross validation is an alternative unbiased model selection option, which can be employed for the proposed model with less combinations.

Additionally, it is important to mention that the following settings were used for all runs: number of annealing steps $N_{\text{anneal}}^{\text{FEM}} = 30$ and the optimization tolerance value $N_{\tau\text{ol}}^{\text{FEM}} = 0.0000000001$.

5.3.2 *Interpretation of optimal model parameters*

For the selection of the model that is optimal with respect to quality and complexity, the introduced mAICc given in Equation (3.70) is deployed. The corresponding optimal number of regimes is denoted N_K^*, analogously the optimal maximal number of transitions is referred to as N_C^*, the optimal memory depth N_M^* of a model, and the optimal choice of explicit external factors is denoted $u^*(t,j)$. Considering all 7938 models, the lowest mAICc value corresponds to a memory-less model (i.e., $N_M^* = 0$) with $N_C^* = 70$, $N_K^* = 3$, and

$$u^*(t,j) = \begin{bmatrix} \text{neigh}_{\text{ice}} \\ \text{temp} \\ \text{CO}_2 \end{bmatrix}. \tag{5.7}$$

The first model matrices of each of the three regimes are inferred to be

$$P_0^{1*} = \begin{bmatrix} 0.1795 & 0.8205 \\ 0.1795 & 0.8205 \end{bmatrix}, P_0^{2*} = \begin{bmatrix} 0.6293 & 0.3707 \\ 0.6293 & 0.3707 \end{bmatrix}, P_0^{3*} = \begin{bmatrix} 0.9416 & 0.0584 \\ 0.9416 & 0.0584 \end{bmatrix}. \tag{5.8}$$

The statistical influence (that is revealed by the model matrices P_e^{1*} for $e \in \{1, \ldots, N_E\}$) of the explicit external factors is discussed separately in Subsection 5.3.3, and the corresponding matrices are given in Appendix C.3 (see (C.1), (C.2), and (C.3)).

The corresponding regime affiliation process $\Gamma^*(t,j)$ is visualized in the panels of Figure 18. The 3D graph in the upper panel shows the affiliations for $t \in \{1, \ldots, N_{T_{\text{train}}}\}$, i.e., starting in January 1989 going to December 2003,

and for the considered arctic area. For visualization reasons, the Hudson Bay area is not displayed and the locations are shown on a quadrangle rather than on a hexagonal grid. The red cells correspond to areas not covered by the data, e.g., locations predominately on one of the continental land masses. The other colors are associated with the three regimes. More precisely, locations j, assigned to local model $P^{1*}(u(t,j))$, i.e, $\gamma_1^*(t,j) = 1$, for fixed t, are displayed in turquoise. Further, cells associated with regime $P^{2*}(u(t,j))$, i.e, $\gamma_2^*(t,j) = 1$, are colored yellow, dark blue relates to $\gamma_3^*(t,j) = 1$, i.e., for fixed time t, locations j can be described best with the transition matrix $P^{3*}(u(t,j))$.

The three optimal local regimes represented by P_0^{k*} and the corresponding affiliations assigned for every time step and each location allow to view the underlying ice dynamics from two different angles. Firstly, it is possible to relate the local models to the two seasonal extremes (i.e., in the broadest sense summer and winter) and an interim phase. These periodic time-wise changes can be seen in both graphs of Figure 18. In the 3D plot in the lower panel of Figure 18 only a short time interval is shown, i.e., $t \in \{1, \ldots, 48\}$, which corresponds to January 1989 to December 1990. Thus, the seasonal regime changes can be seen better. Although the time-wise change seems to effect the regime assignments, the main dependency is related to the geographical regions (e.g., associated to parallels with a certain latitude). In other words, the development in time of the corresponding optimal model parameter $\Gamma^*(t,j)$ strongly depends on the location.

For instance, as expected, locations in the pole region usually are affiliated with a probability close to one to be in the state of ice the entire year (i.e., $P^{3*}(u(t,j))$), whereas locations near the marginal seas of the arctic ocean (e.g., Barents Sea or Chukchi Sea) do react to seasonal change but have high probabilities to be in the liquid state (i.e., $P^{1*}(u(t,j))$) even in the winter months.

Concluding, the process Γ^* also identifies thicker multi-year ice sheets, which are less prone to decrease due to their internally stronger structure and their geographical location, which might not permit sea ice loss caused by shelfs drifting (accelerated by ocean currents or wind) in the open sea.

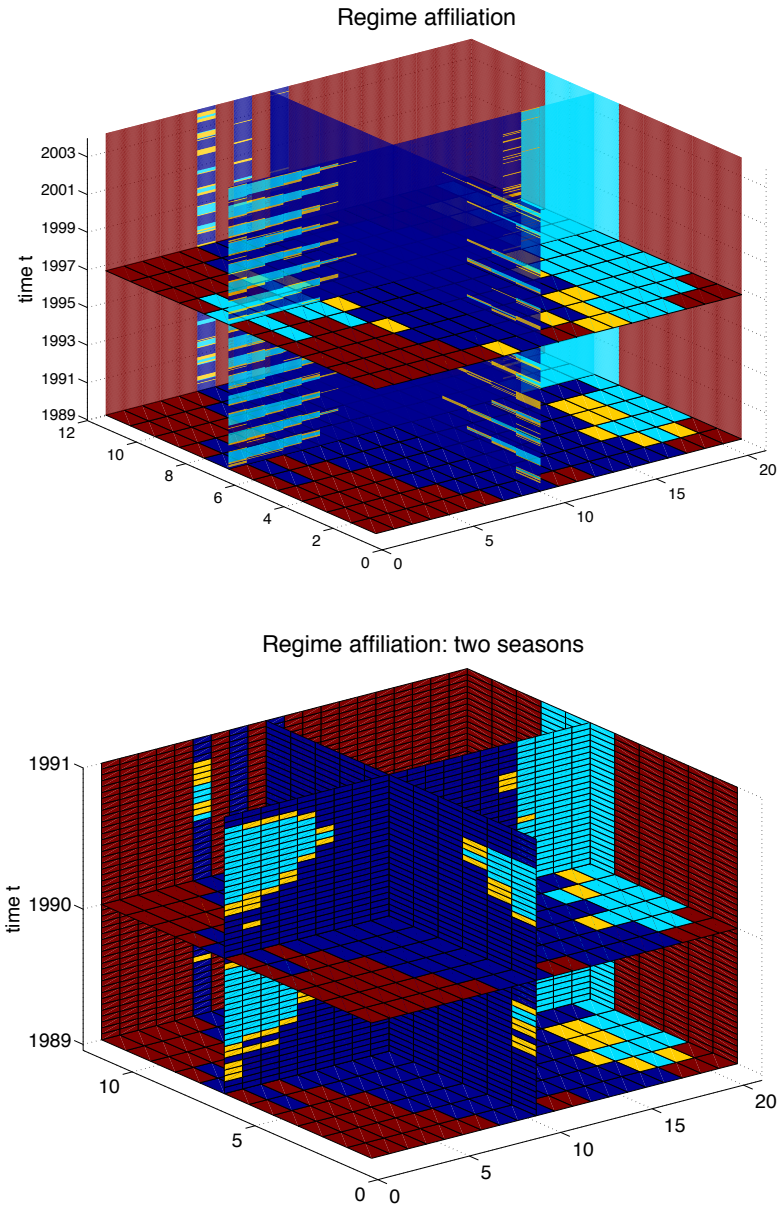

Figure 18: *The figure shows a 3D visualization of the regime affiliations $\gamma_k(t,j)$ for different locations on two time scales. More precisely, the graph on the upper is plotted from 1989 to 2004, where the graph on lower depicts a time interval from 1989 to 1991. The three regimes are associated with the following colors: turquoise relates to regime 1, i.e, $\gamma_1^*(t,j) = 1$, yellow corresponds to regime 2, i.e, $\gamma_2^*(t,j) = 1$, and for $\gamma_3^*(t,j) = 1$ locations are colored dark blue. Further, red is associated with empty cells (i.e., no data available), which do not correspond to any regime and thus have a structure similar to the surrounding continental land masses. For reasons related to the visualization restraints, the locations are considered on a rectangular lattice instead of a hexagonal grid. Note that the affiliations of locations in the Hudson Bay are not displayed.*

5.3.3 *Statistical impact*

The summer minimum of the ice extent has decreased considerably in the past decade and an ice free arctic[11] is a likely climate scenario to occur this century [58, 118]. Various factors such as changes in the atmospheric circulation or the brine structure of the ice are associated with this retreat. Yet, there is still discordancy among scientists over the influence magnitude of the different quantities on the diminishing sea ice. Thus, one aim is to estimate the statistical impact of the used explicit external factors on the basis of the determined model to understand their influence on the sea ice dynamics. The individual impact of each explicit external factor $u_e(t, j)$ is considered via the absolute value of the corresponding model matrix entries of each regime. As the considered model

$$\mathcal{M}^{\text{Markov}}(3, 70, 0, [\text{neigh}_{\text{ice}}, \text{temp}, \text{CO}_2]^\top) \tag{5.9}$$

is independent of previous state probabilities, the absolute values of the entries of the model matrices $\{P_e^k\}_{mn}$ are equal for all $m, n \in \{1, \ldots, N_S\}$, for fixed $e \in \{1, \ldots, N_E\}$, and fixed $k \in \{1, \ldots, N_k\}$. Thus, for a fixed regime k and a specific explicit external factor index e, it is possible to associate the corresponding statistical impact $\mathcal{I}(e, k)$ of the factor on the state of the underlying process with the absolute value of one matrix entry of $\{P_e^k\}$, i.e.,

$$\mathcal{I}(e, k) = \left\{P_e^{k*}\right\}_{11} \quad \text{for } e \in \{1, \ldots, N_E\} \text{ and } k \in \{1, \ldots, N_k\}. \tag{5.10}$$

Further, the relative statistical impact

$$\mathcal{I}_{\text{rel}}(e, k) = \frac{\mathcal{I}(e, k)}{\sum_{r=1}^{N_E} \mathcal{I}(r, k)} \quad \forall\, e, k \tag{5.11}$$

is considered. Due to the previous scaling of the explicit external factors to the $[-1, 1]$ interval, the absolute values are not related to any scales corresponding to the respective quantity and its unit. The results corresponding to the considered optimal model $\mathcal{M}^{\text{Markov}}(3, 70, 0, [\text{neigh}_{\text{ice}}, \text{temp}, \text{CO}_2]^\top)$ are presented in Table 1.

In general, it can be said that the statistical results show strong existing spatial correlations. In particular, it is revealed that the neighbors are the main influencing component in the context of a high probability for $\sigma(t, j, l)$

11 Commonly, a sea ice extent of only $1,000,000$ km^2 is defined as an *ice free* arctic.

to be in liquid state respectively to change into it (see (5.8)). Furthermore, the previous state probabilities of the neighbors are predominately relevant for the local model representing an intermediate phase, i.e., $P^{2*}(u(t,j))$. However, the temperature has the strongest relative impact in areas respectively time intervals affiliated with the optimal model matrix $P^{2*}(u(t,j))$. Concluding, although CO_2 does influence the state of microscopic sea ice in the arctic, an immediate negative trend is procrastinated by the rather small determined statistical impact. The influence of changing temperature values on the other hand directly effects existing multiyear ice sheets (which are mostly characterized by the local model matrix $P^{3*}(u(t,j))$).

$u_e(t,j)$	$\mathcal{I}_{\text{rel}}(e,1)$	$\mathcal{I}(e,1)$	$\mathcal{I}_{\text{rel}}(e,2)$	$\mathcal{I}(e,2)$	$\mathcal{I}_{\text{rel}}(e,3)$	$\mathcal{I}(e,3)$
$\text{neigh}_{\text{ice}}$	100%	0.1795	68.52%	0.2002	39.01%	0.0130
temp	0%	$0.2 \cdot 10^{-16}$	30.3%	0.0887	45.53%	0.0152
CO_2	0%	$0.1 \cdot 10^{-16}$	1.1%	0.0032	15.46%	0.0051

Table 1: *Shows the absolute (see* (5.10)*) and relative statistical impact (see* (5.11)*) of the three explicit external factors: neighboring ice, temperature, and CO_2.*

5.3.4 *Out-of-sample-performance*

In the following, the out-of-sample performance of the model

$$\mathcal{M}^{\text{Markov}}(3, 70, 0, [\text{neigh}_{\text{ice}}, \text{temp}, \text{CO}_2]^\top) \tag{5.12}$$

is examined. In pursuance of estimating the affiliations outside of the training set $\{1, \ldots, 360\}$, a stationary, homogenous model $\mathcal{M}^{f^\Gamma,\Gamma}(1, -, N_M^\Gamma, u^\Gamma(t,j))$ for $f^\Gamma \in \{\text{Markov}, \text{logit}\}$ is fitted to Γ^*. As has been illustrated in Subsection 5.3.2, the affiliations $\gamma_k^*(t,j)$ exhibit periodic behavior for certain locations j. Due to the fact that the model describing these affiliations is assumed to be stationary and homogenous, it is sensible to consider external factors $u_e^\Gamma(t,j)$ with similar periodic oscillations. These additional external factors are computed by means of Γ^* (see calculation in Algorithm 8). Subsequently, the set of explicit external factors considered for the parametrization of Γ^* is defined as

$$\mathcal{E}^\Gamma := \{\text{neigh}_{\text{ice}}, \text{neigh}_{\text{land}}, \text{temp}, \text{CO}_2, \text{NAO}, \text{AO}, \text{period}_1, \text{period}_3\}. \tag{5.13}$$

Note that all explicit external factors contained in $\mathcal{E}$ are used for the inference of the dynamics of the regime assigning process. In other words, not all

combinations are considered. Yet, different models $\mathcal{M}^{f^\Gamma,\Gamma}(1,-,N_M^\Gamma,u^\Gamma(t,j))$ with $u^\Gamma(t,j) \in \mathcal{E}^\Gamma$ are computed with respect to the linear dependency on the additional external factors period_k given in Line 8 of Algorithm 8.

Algorithm 8: Periodic behavior of Γ^*

input : Γ^*
output: period_k for $k \in \{1,\dots,N_K^*\}$

```
1  N_Y = N_{T_train}/24
2  for k = 1 : N_K^* do
3      for j = 1 : N_J do
4          S = 0
5          for y = 1 : N_Y do
6              r = (y − 1)24 + 1
7              s = (y − 1)24 + 24
8              S = S + γ_k(r : s, j)
9          for y = 1 : N_Y + 1 do
10             r = (y − 1)24 + 1
11             s = (y − 1)24 + 24
12             period_k(r : s, j) = round (S/15)
```

The corresponding mAICc values can be found in Table 6 in Appendix C.3. The lowest value is attained for the model

$$\mathcal{M}^{\text{logit},\Gamma}(1,-,0,[\text{neigh}_{\text{ice}},\text{neigh}_{\text{land}},\text{temp},\text{CO}_2,\text{NAO},\text{AO},\text{period}_1]). \quad (5.14)$$

Estimates of the state probabilities $\pi(t,j)$ for $t > N_{T_{\text{train}}}$ can thus be computed via Algorithm 7. The resulting approximations for 4 different example locations (for geographical information on the locations see graphic on the right in Figure 19) are visualized in Figures 20 and 21. The actual data is also displayed as a reference and the start of the out-of-sample prediction is marked with a vertical red line.

Although the ice evolutions of the four example locations are of very different nature, the approximations have a very high quality. In particular, the performance of the considered model for $t \in \{N_{T_{\text{train}}}+1,\dots,N_T\}$ is promising. In Figure 22 the predictive capability of the model is considered for all locations j for a fixed time step $N_{T_{\text{train}}}+5$ (which corresponds to March 2004). In detail a visualization of the rounded data and the approximation are displayed. Further, land cells are added to the image in order to show the continental structure of the arctic circle. Again, the accuracy of the approximation is very high.

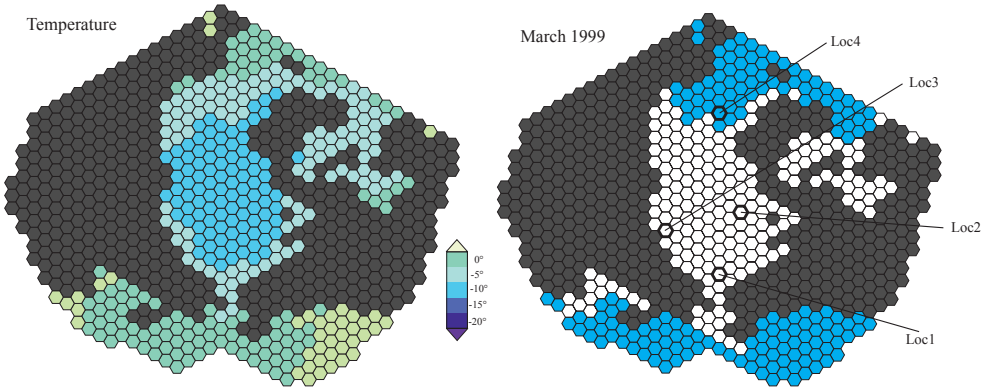

Figure 19: *The image on the shows the rounded data $\pi_1(t, j)$ observed in March 1999 (i.e., $t = 269$). The geographic positions of four example locations are shown.*

Concluding, it is possible to characterize the complex dynamics underlying the arctic sea ice coverage data with the obtained model. Summarizing, again this result emphasizes the fact that the introduced framework allows to infer suitable models even though only a small portion of the relevant information is given. In particular, it is important to note that the considered system is highly complex and the corresponding data is multidimensional, which makes the identification of the underlying process even more difficult.

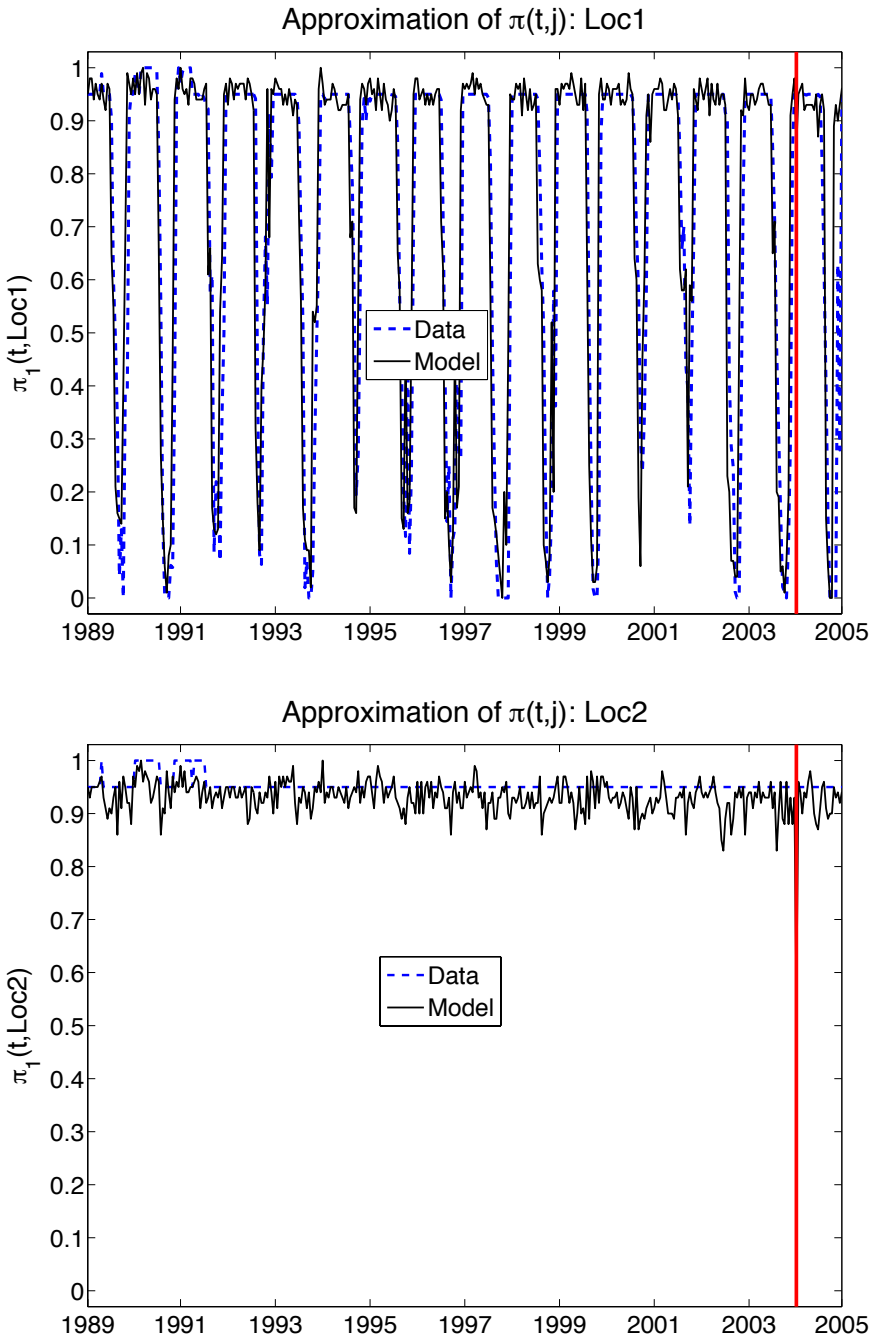

Figure 20: *The graphics show the approximations of $\pi_1(t, Loc1)$ and $\pi_1(t, Loc2)$ (Loc1 and Loc2 are defined in the graphic on the right in Figure 19) for $t \in [1989, 2005]$. In particular, the prediction of 24 time steps (January 2004 to December 2004) is displayed (the start is marked by a vertical red line). Additionally, the actual data is given as a reference.*

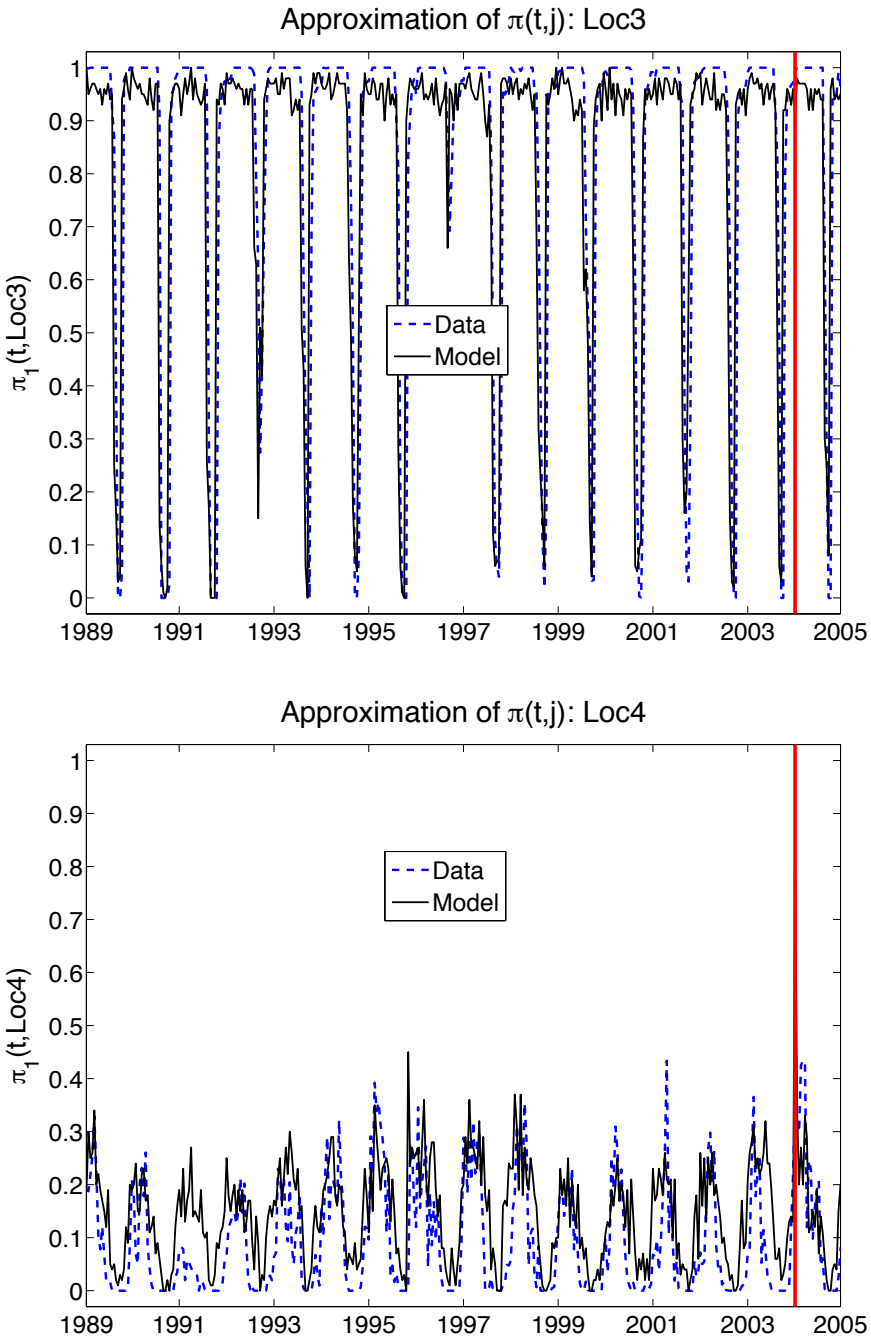

Figure 21: *Comparisons of the observed data $\pi_1(t,j)$ for $t \in \{1,\dots N_T\}$ and approximations of it are shown for two different locations: Loc3 and Loc4 (see graphic on the right in Figure 19). Note that the time step $N_{T_{train}}$ is emphasized via a red vertical line.*

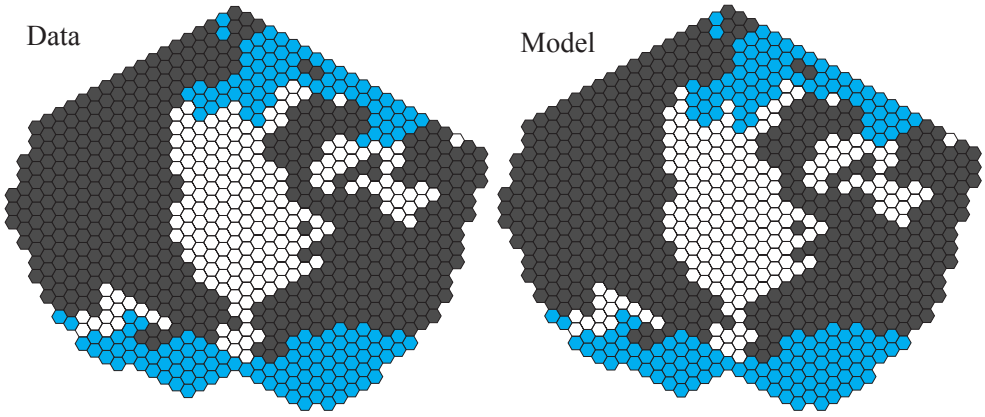

Figure 22: *The rounded data $\pi_1(N_{T_{rain}} + 5, j)$, i.e., observed in March 2004, is shown for all j on the upper panel, and on the lower panel the corresponding approximation determined with the model $\mathcal{M}^{Markov}(3, 70, 0, [neigh_{ice}, temp, CO_2]^\top)$ is displayed.*

SUMMARY

In this thesis, a spatial extension of the existing non-stationary Markov regression was developed. The aim was to introduce a framework that allows to characterize spatio-temporal Markov processes with a finite state space governed by external influences. This was achieved by extending the available structure of the purely time-dependent model. Concluding, the presented framework is bridging the gap between the considered state-of the-art data-based analysis tool and the wide range of dynamical systems with underlying spatio-temporal processes.

While standard modeling frameworks such as SVMs and ANNs lack the ability to take unresolved external factors into account, the proposed model integrates these implicit quantities via explicitly time- and space-dependent model parameters. More precisely, in the presence of unresolved external factors the derived Markov model has been theoretically verified to have a non-stationary, non-homogenous expression similar to the non-stationary model structure in the purely time-wise case [53]. The numerical approach to fit a non-stationary (i.e., purely time-dependent) Markov model on the basis of available data has been realized via gradient-based optimization of a corresponding regularized inverse problem [84]. Following this idea, an algorithm to compute a model with an additional spatial dependency was outlined and implemented.

As optimization via gradient-based approaches, which are standardly employed in the context of clustering algorithms, does not necessarily provide global optimal solutions to the posed inverse problem, the theoretical and numerical aspects of using an alternative minimization approach were presented. More specifically, a coupling of the considered non-stationary, non-homogenous Markov regression framework with an MCMC-based minimization ansatz was developed for the computation of the optimal model parameters. In contrast to the standardly employed optimization tools, this MCMC approach allows with high probability to compute a global minimizer and reduces the computational complexity considerably.

An artificial dynamical systems was used to experimentally verify the theoretically derived properties of the model. In particular, the capabilities of the developed model to accurately describe dynamical processes that are predominately influenced by implicit quantities were successfully demon-

strated [30, 31, 53]. While data estimates computed with non-dynamical approaches, e.g., ANN, are distorted for such systems, the approximation determined with the proposed non-stationary, non-homogenous model is very good.

Due to the recent rapid decline of sea ice, a lot of research has been focused on understanding the underlying dynamics. Here a data-based approach was used to gain a new perspective. In detail, the proposed Markov regression framework was employed to infer a model describing the evolution of aggregate states of water molecules in the arctic ocean. This model was interpreted to gain more information on the involved interactions. In particular, it became apparent that there are strong correlations between neighboring locations. Further, the evolution of the aggregate states could be linked to an explicitly time- and space-dependent affiliation process. This means that the underlying system is evidently influenced by unresolved quantities.

In the context of climate applications such as the problem of understanding the arctic sea ice variability, there are many important open questions, some of which will be stated now:

- What are the time-lags with respect to the impact of certain explicit external factors?
- Can the presented model replace a regional model (e.g., a regional ice model) that is coupled to a global climate model (e.g., an atmospheric model)? And how can the corresponding interactions between these models be realized?
- Can the computational complexity of the framework be improved so that even bigger data sets can be considered? And to what degree can parallel computing reduce the run time?
- Is an additional upper bound, restricting the number of transitions of the affiliation process with respect to the locations, beneficial for the interpretation of the model? And how can such an additional constraint be made computationally feasible?
- How can the models be used to examine phase transitions? And what are the limitations in the context of studying certain responses triggered by tuned explicit external factors with respect to the model and the vector space of explicit external factors?

The tools presented in this thesis are an essential step in the direction of answering these questions.

ZUSAMMENFASSUNG

Prozesse in natürlichen dynamischen Systemen mit skalenübergreifenden Wechselwirkungen können oft nicht zufriedenstellend mit rein deterministischen Modellen beschrieben werden. Daher ist eine stochastische Beschreibung in vielen Fällen eine gute Alternative, um diese Wechselwirkungen zu simulieren. Stochastische Prozesse werden häufig in vier Klassen unterteilt, abhängig von der Kardinalität ihres Zustandsraums, abzählbar oder überabzählbar, und ihrer zeitlichen Entwicklung, diskret oder stetig. Das bekannteste Beispiel für die hier betrachteten zeitdiskreten stochastischen Prozesse sind die Markov-Prozesse. Diese Prozesse beschreiben eine Dynamik, die nur von ihrem vorherigen Zustand abhängt. In der Realität werden diese Prozesse häufig von äußeren Faktoren angetrieben. Da ein Standard-Markov-Prozess keine direkte Modellierung dieser treibenden Einflüsse erlaubt, wurde ein Markov-Modell mit einer speziellen Struktur, die es erlaubt, diese externen Faktoren linear in die Modellparameter einfließen zu lassen, von Illia Horenko entwickelt. Zusätzlich wurde von ihm gezeigt, dass Faktoren, zu denen kein direkter Zugang besteht, durch eine explizite Zeitabhängigkeit der Modellparameter beschrieben werden können. Anhand von Beobachtungsdaten ist es möglich, diese Modellparameter zu approximieren. Eine zugehörige Methode, die diese Art von Modellierung von diskreten Prozessen mit nicht-stationären Modellparametern erlaubt, wird FEM-BV-Clustering-Ansatz genannt und wurde ebenfalls von Illia Horenko entwickelt.

Da die meisten Systeme nicht auf eine rein zeitliche Entwicklung beschränkt sind, sondern auch eine räumliche Dynamik aufweisen, wurde in dieser Dissertation eine räumliche Erweiterung des Markov-Modells und der zugehörigen FEM-BV-Clustering-Methode entwickelt und getestet.

Dabei wurde gezeigt, dass, wie schon in der rein zeitlichen Beschreibung, jeglicher Einfluss impliziter (sprich unaufgelöster) Faktoren durch eine explizit zeitliche und nun auch räumliche Abhängigkeit ausgedrückt werden kann. Diese theoretische Eigenschaft des Modells wurde experimentell anhand von künstlichen Testsystemen überprüft. Hierfür wurde der bestehende FEM-BV-Clustering-Algorithmus erweitert, um auch räumliche Modelle bestimmen zu können.

In diesem Zusammenhang wurde ein MCMC-basierter Optimierungsalgorithmus als Alternative zu den bisher genutzten Standardverfahren der linearen Optimierung (z.B. Simplex-Verfahren) in den FEM-BV-Clustering-Algorithmus eingebettet. Die Vorteile dieser MCMC-Methode sind, dass die Laufzeit für bestimmte Beispiele deutlich reduziert werden kann und dass der Algorithmus ein globales Minimum liefert.

Um die betrachteten Algorithmen auf künstliche und reale Datensätze anwenden zu können, wurden sie im Rahmen dieser Dissertation in der Programmiersprache C++ implementiert.

Im direkten Vergleich mit Standardmethoden schneiden die neuen Me-thoden für die gewählten Testsysteme sehr gut ab. Insbesondere für den betrachteten Prozess mit vielen unaufgelösten Faktoren, an dem die benutzten Standardmethoden (d.h. Support Vector Machines und Künstliche neuronale Netze) scheitern, ist das erweiterte Markov-Modell überlegen.

Weiterhin wurde die räumliche Erweiterung für die Modellierung eines durch Eisbedeckungsdaten der Arktis gegebenen Raum-Zeit-Prozesses genutzt. Wegen des starken Rückgangs des Meereises im Arktischen Ozean und der damit verbundenen negativen Konsequenzen ist es besonders wichtig, alle Aspekte dieses komplexen dynamischen Systems besser zu verstehen. Ein unter physikalischen Gesichtspunkten unvoreingenommener datenbasierter Ansatz kann zudem neues Licht auf Aspekte werfen, die durch die üblichen Methoden (z.B. Klimamodelle) nicht zum Vorschein gebracht werden können.

Das in dieser Arbeit entwickelte erweiterte Markov-Modell wurde genutzt, um die räumliche und zeitliche Entwicklung von Aggregatzuständen von Wassermolekülen in der Arktis zu beschreiben. Mit Hilfe des errechneten Modells konnten qualitativ hochwertige Simulationen der Daten erzeugt werden. Außerdem wurden statistische Einflusswerte für alle involvierten expliziten Faktoren bestimmt. Insbesondere sind sowohl der Einfluss der Nachbarschafts-Konstellation als auch die starke Abhängigkeit des unterliegenden Prozesses von den unaufgelösten Faktoren deutlich geworden.

A

STEP 2 OF SUBSPACE ALGORITHM

The subspace algorithm [31, 52, 84], described in Section 3.5, is an iteration over two optimization steps. In the following section the numerical details of optimization Step 2 of Algorithm 1 for the considered Markov model (see (3.22)) are outlined including a discussion on the necessary implementation steps. The corresponding commented source code, implemented in C++ can be found on `http://www.dewiljes.de/dewiljes/Jana.html`. For the computations in Chapter 5, a C++ code of Step 1, implemented by Philipp Metzner, was deployed.

A.1 NON-STATIONARY NON-HOMOGENOUS MARKOV REGRESSION

The problem of minimizing the functional $\mathbf{L}(\Gamma(t,j), P(u(t,j)))$, given in (3.34), for fixed $\Gamma(t,j)$ with respect to model parameters $P(u(t,j))$, subject to linear constraints (3.37), (3.38), and (3.41) (and optional (3.42)), belongs to the class of *Quadratic programming* problems [42]. These mathematical optimization problems can be expressed as follows:

$$\mathbf{L}(\mathbf{p}) = 0.5\mathbf{p}^\top \mathbf{G}\ \mathbf{p} + \mathbf{g0}^\top \mathbf{p} \rightarrow \min_{\mathbf{p}}. \tag{A.1}$$

Before applying a quadratic programming solver[1] it is necessary to bring the regarded problem (3.34) in the particular form given in (A.1). Due to the fact that a set of minimal model matrices $\{P_0^k, \ldots, P_{N_E}^k\}$ for a fixed affiliation process $\Gamma(t,j)$ can be determined independently for each $k \in \{1, \ldots, N_K\}$,

1 For the computations in Chapter 5 an open source solver, that can be found on `http://www.diegm.uniud.it/digaspero/index.php?page=software`, was employed. For a detailed information on the algorithm the reader is referred to [42].

the derivation is done for the stationary, homogenous case, i.e., $N_K = 1$. Consequently, the aim is to determine the model parameter function

$$P(t, j, u(t, j)) = \sum_{e=0}^{N_E} u_e(t, j) P_e, \tag{A.2}$$

i.e., to find model matrices P_e for $e \in \{0, \ldots, N_E\}$. Note that $u_0(t, j) := 1$ for all t and j and is an artificial entry in the vector of external factors which is used to make the following vector notation more comprehensive. For the regarded Markov model choice (see (3.22)) the unknown model matrices can be expressed in vector form by assembling the columns of the $N_E + 1$ matrices P_e, i.e.,

$$\mathbf{p} = (\mathbf{vec}(P_0), \ldots, \mathbf{vec}(P_{N_E})) \in \mathbb{R}^{(N_E+1)N_S^2} \tag{A.3}$$

with

$$\mathbf{vec}(P_e) = (P_e(\cdot, 1), \ldots, P_e(\cdot, n)) \in \mathbb{R}^{N_S^2} \quad e \in \{0, \ldots, N_E\}. \tag{A.4}$$

To define the corresponding symmetric matrix $\mathbf{G}$, the functional is separated into its quadratic and its linear component:

$$\begin{aligned}
&L(P(u(t,j))) \\
&\quad = \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} ||\pi(t+1,j)^T - \pi(t,j)^T P(u(t,j))||_2^2 && \text{(A.5)} \\
&\quad = \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} \Big\langle \pi(t+1,j)^T - \pi(t,j)^T P(u(t,j)), \\
&\qquad\qquad \pi(t+1,j)^T - \pi(t,j)^T P(u(t,j)) \Big\rangle && \text{(A.6)} \\
&\quad = \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} \Big(\Big\langle \pi(t+1,j)^T, \pi(t+1,j)^T \Big\rangle - 2 \Big\langle \pi(t+1,j)^T, \pi(t,j)^T P(u(t,j)) \Big\rangle \\
&\qquad\qquad + \Big\langle \pi(t,j)^T P(u(t,j)), \pi(t,j)^T P(u(t,j)) \Big\rangle \Big). && \text{(A.7)}
\end{aligned}$$

The linear part (second summand) of the first part of (A.7) can easily be reshaped to be dependent on $\mathbf{p}$, i.e.,

$$\begin{aligned}
&\left\langle \pi(t+1,j)^T, \pi(t,j)^T P(u(t,j)) \right\rangle \\
&\quad = \sum_{e=0}^{N_E} u_e(t,j) \left\langle \pi(t+1,j)^T, \pi(t,j)^T P_e \right\rangle && \text{(A.8)} \\
&\quad = \sum_{e=0}^{N_E} u_e(t,j) \left\langle \mathbf{vec}(\pi(t,j)^T \pi(t+1,j)^T), \mathbf{vec}(P_e) \right\rangle && \text{(A.9)} \\
&\quad = \langle g0(t,j), \mathbf{p} \rangle && \text{(A.10)}
\end{aligned}$$

with

$$\begin{aligned}
&g0(t,j) && \text{(A.11)} \\
&\quad := \left(u_0(t,j)\mathbf{vec}(\pi(t,j)\pi(t+1,j)^T), \ldots, u_{N_E}(t,j)\mathbf{vec}(\pi(t,j)\pi(t+1,j)^T \right) \\
&\quad \in \mathbb{R}^{(N_E+1)N_S^2 \times 1}.
\end{aligned}$$

To match the notation in (A.1), $\mathbf{g0}$ is defined to be equal to the sum over all locations and time steps multiplied by -2, i.e.,

$$\mathbf{g0} = -2 \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} g0(t,j). \tag{A.12}$$

Further, considering the quadratic component of the term (see (A.7))

$$\begin{aligned}
&\left\langle \pi(t,j)^T P(u(t,j)), \pi(t,j)^T P(u(t,j)) \right\rangle \\
&= \sum_{e_1=0}^{N_E} \sum_{e_2=0}^{N_E} u_{e_1}(t,j) u_{e_1}(t,j) \left\langle \pi(t,j)^T P_{e_1}, \pi(t,j)^T P_{e_2}) \right\rangle && \text{(A.13)} \\
&= \sum_{e_1=0}^{N_E} \sum_{e_2=0}^{N_E} u_{e_1}(t,j) u_{e_1}(t,j) \left\langle \mathbf{vec}(P_{e_1}), \mathbf{diag}(\pi(t,j)\pi(t,j)^T)\mathbf{vec}(P_{e_2})) \right\rangle && \text{(A.14)}
\end{aligned}$$

it is possible to define $G(t,j) \in \mathbb{R}^{(N_E+1)N_S^2 \times (N_E+1)N_S^2}$, constituted of blocks

$$\{G(t,j)\}_{r_1:i_1,r_2:i_2} = u_{e_1}(t,j) u_{e_2}(t,j) \mathbf{diag}(\pi(t,j)\pi(t,j)^T) \in \mathbb{R}^{N_S^2 \times N_S^2} \tag{A.15}$$

with

$$r_1 = e_1 N_S + 1 \tag{A.16}$$

$$r_2 = e_2 N_S + 1 \tag{A.17}$$

$$i_1 = e_1 N_S + N_S \tag{A.18}$$

$$i_2 = e_2 N_S + N_S \tag{A.19}$$

for $e_1, e_2 \in \{0, \ldots, N_E\}$. Then the matrix $\mathbf{G}$ can be expressed as follows:

$$\mathbf{G} = 2 \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} G(t, j). \tag{A.20}$$

Note that the remaining first summand in (A.7) is not dependent on $\mathbf{p}$ and, therefore, is not relevant for the optimization and consequently will not be regarded. Concluding, the original problem given in (3.34) can be written in the terminology used for mathematical optimization problems (see Eq. (A.1)). More specifically, as already mentioned above, the quadratic optimization problem has to be solved independently for each $k \in \{1, \ldots, N_K\}$ for fixed regime assignments $\Gamma(t, j)$, i.e.,

$$\mathbf{L}(\mathbf{p}^k) = 0.5(\mathbf{p}^k)^\top \mathbf{G}^k \mathbf{p}^k + (\mathbf{g0}^k)^\top \mathbf{p}^k \rightarrow \min_{\mathbf{p}^k} \ \forall k, \tag{A.21}$$

where

$$\mathbf{G}^k = 2 \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} \gamma_k(t, j) G(t, j) \tag{A.22}$$

and

$$\mathbf{g0}^k = -2 \sum_{t=1}^{N_T} \sum_{j=1}^{N_J} \gamma_k(t, j) g0(t, j). \tag{A.23}$$

Further, each of the N_K quadratic optimizations problems, given in (A.21), is subject to linear inequality and equality constraints. The details are discussed in the following.

A.1.1 *Constraints*

The inequality and equality constraints (3.37), (3.38), and (3.41) are reformulated to be subject to the vector $\mathbf{p}$. The constraints are employed to ensure that the model parameter $P(t, j, u(t, j))$ is a stochastic matrix for all t, j

and $u(t,j)$. The equality constraints (3.37) and (3.38) are combined in the following equation:

$$\mathbf{CE}(N_M)^\top \cdot \mathbf{p} + \mathbf{ce0} = \mathbf{0} \tag{A.24}$$

with

$$\mathbf{CE}(1) := \underbrace{\begin{bmatrix} \mathcal{Q}(\mathbf{Id}_{N_S}) & 0 & \dots & 0 \\ 0 & \mathcal{Q}(\mathbf{Id}_{N_S}) & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \dots & 0 & \mathcal{Q}(\mathbf{Id}_{N_S}) \end{bmatrix}}_{\in \mathbb{R}^{(N_E+1)N_S^2 \times (N_E+1)N_S}} \tag{A.25}$$

and

$$\mathbf{ce0} := \begin{bmatrix} -\mathbf{1}_{N_S} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{(N_E+1)N_S \times 1}, \tag{A.26}$$

where the auxiliary matrix is defined as follows:

$$\mathcal{Q}(\mathbf{Id}_{N_S}) = \underbrace{\begin{bmatrix} \mathbf{Id}_{N_S} \\ \mathbf{Id}_{N_S} \\ \vdots \\ \mathbf{Id}_{N_S} \end{bmatrix}}_{\in \mathbb{R}^{N_S^2 \times N_S}}. \tag{A.27}$$

Additionally to the general model, defined in (3.22), the independent special case of the Markov model is considered (see discussion in Subsection 3.2.1). Essentially, a memoryless process is assumed (i.e., $N_M = 0$). This independence of the current state probabilities is realized by ensuring that the columns of P_e^k have equal entries (see Equation (3.42)). Thus, the matrix **CE** has to be enhanced so that this additional constraint is fulfilled:

$$\mathbf{CE}(0) = \underbrace{\begin{bmatrix} \mathcal{Q}(\mathbf{Id}_{N_S}) & 0 & \dots & 0 & \mathcal{A} & 0 & \dots & 0 & 0 \\ 0 & \vdots & \dots & 0 & 0 & \mathcal{A} & \dots & 0 & 0 \\ 0 & \mathcal{Q}(\mathbf{Id}_{N_S}) & \dots & \vdots & 0 & \ddots & \ddots & 0 & 0 \\ \vdots & 0 & \ddots & 0 & 0 & \dots & 0 & \mathcal{A} & 0 \\ \vdots & \vdots & & 0 & 0 & \dots & 0 & 0 & \mathcal{A} \\ 0 & 0 & 0 & \mathcal{Q}(\mathbf{Id}_{N_S}) & 0 & \dots & 0 & 0 & 0 \end{bmatrix}}_{\in \mathbb{R}^{(N_E+1)N_S^2 \times (N_E+1)N_S + (N_E+1)(N_S-1)(N_S-1)}} \tag{A.28}$$

with

$$\mathbf{ce0} = \begin{bmatrix} -\mathbf{1}_{N_S} \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix} \in \mathbb{R}^{(N_E+1)N_S+(N_E+1)(N_S-1)(N_S-1)\times 1}, \tag{A.29}$$

where the corresponding auxiliary matrix is defined as follows:

$$\mathcal{A} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -1 & \ddots & 0 \\ 0 & \ddots & 1 \\ 0 & 0 & -1 \end{bmatrix}}_{\in \mathbb{R}^{N_S\times(N_S-1)}}. \tag{A.30}$$

Note that one can switch between a model with and without memory by simply choosing the associated matrix **CE**, i.e., (A.25) for $N_M = 1$ or (A.28) for $N = M = 0$, for the optimization procedure. In order to ensure the required inequality Equation (3.41), the following expression is used with the formulation:

$$\mathbf{CI}^\top \cdot \mathbf{p} + \mathbf{ci0} \geq \mathbf{0}. \tag{A.31}$$

As inequality (3.40) does not depend on the vector of explicit external factors $u(t, j)$, the formulation is straightforward

$$\underbrace{(\mathbf{Id}_{N_S^2}, \mathbf{0}, \ldots, \mathbf{0})\mathbf{p} \geq \mathbf{0}}_{\in \mathbb{R}^{N_S^2\times(N_E+1)N_S^2}}. \tag{A.32}$$

As already discussed in Section 3.2, the original non-negativity requirement of the entries of the model matrices $P(u(t, j))$ (see (3.39)) mainly depends on $u(t, j)$ and, therefore, in general is not computationally feasible. Yet, it is possible to reduce the number of necessary inequalities by assuming that the convex-hull of the set $\mathcal{U}$, containing vector $u(t, j)$, is a N_E-dimensional hypercube

$$\mathcal{H}_{Cube} = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_m, b_m]. \tag{A.33}$$

As the minima and maxima of a hypercube are its corners, i.e.,

$$a_e = \min\{u_e(t, j) : j = 1, \ldots, N_J \text{ and } t = 1, \ldots, N_T\} \tag{A.34}$$

and

$$b_e = \max\{u_e(t,j) : j = 1,\dots,N_J \text{ and } t = 1,\dots,N_T\}, \tag{A.35}$$

it is possible to reduce the dimensionality of the original constraint (see (3.39)) to the simpler numerical feasible inequality given in (3.41). In detail, all 2^{N_E} combinations

$$\{a_1,b_1\} \times \{a_2,b_2\} \times \cdots \times \{a_{N_E},b_{N_E}\} \tag{A.36}$$

need to be checked. Thus, leading to the following definition of the matrix describing the inequalities:

$$\mathbf{CI} = \underbrace{\begin{bmatrix} \mathbf{Id}_{N_S^2} & \dots & \mathbf{Id}_{N_S^2} & \dots & \mathbf{Id}_{N_S^2} \\ u_1^{\text{comb}(1)}\mathbf{Id}_{N_S^2} & \dots & \dots & u_1^{\text{comb}(2^{N_E})}\mathbf{Id}_{N_S^2} & \mathbf{0} \\ u_2^{\text{comb}(1)}\mathbf{Id}_{N_S^2} & \dots & \dots & \vdots & \mathbf{0} \\ \vdots & \dots & \dots & \vdots & \vdots \\ \vdots & \dots & \dots & u_{N_E-1}^{\text{comb}(2^{N_E})}\mathbf{Id}_{N_S^2} & \vdots \\ u_{N_E}^{\text{comb}(1)}\mathbf{Id}_{N_S^2} & \dots & \dots & u_{N_E}^{\text{comb}(2^{N_E})}\mathbf{Id}_{N_S^2} & \mathbf{0} \end{bmatrix}}_{\in \mathbb{R}^{(N_E+1)N_S^2 \times (2^{N_E}+1)N_S N_S}} \tag{A.37}$$

with $u_e^{\text{comb}(r)}$ taking values in $\{a_e,b_e\}$ according to the current combination $r \in \{1,\dots,2^{N_E}\}$. Note that the matrix component of **CI**, given in (A.32), is already included.

B

ADAPTIVE SIMULATED ANNEALING SCHEME

In Subsection 3.5.1 a stochastic optimization approach for the sub-optimizations $\mathbf{L}_j^{\mathfrak{r}}(\Gamma(:,j))$ (see (3.46)) of the Tikhonov-regularized inverse problem was proposed. The ansatz is based on generating samples according to a particular Boltzmann distribution (see (3.51)) employing a RWM algorithm. Additionally, a complementary adaptive tuning method of the inverse temperature variable β and of the noise factor $\mathfrak{n}$ during the Metropolis algorithm run has been proposed [30, 37]. The acceptance/rejection-procedure has already been considered and specified in Algorithm 2, thus, the subsequent explanations focus on the simulated annealing scheme. In particular, the corresponding pseudocode of the methodology is outlined in detail.

The key idea of the presented framework is to increase β at a sensible pace, ensuring enough flexibility to traverse the sample space and, at the same time, keeping the rate of accepted samples at an optimal level, which is theoretically verified to be $23,4\%$ [90], by adaptively changing the noise factor $\mathfrak{n}$.

The update of the noise factor $\mathfrak{n}$ in Lines 9-15 is set to take place after 1000 iterations (see Line 9) of the acceptance-procedure (see Lines 3-8). The chosen frequency allows to give a good statistical overview of the total number of accepted samples N_{accept} while giving the regular opportunity to make necessary changes. Nevertheless, the suitability of the frequency value should be tested for every application or different energy function. In pursuance of keeping the ratio of accepted samples and the total number of proposed samples in the range of the reference value $23,4\%$, an interval, going from 18% (see line 10: $0.18 = \frac{90}{500}$) to 28% (see line 12: $0.28 = \frac{140}{500}$), is considered. In case the number of accepted samples is outside this considered interval of percentages, the noise factor is adaptively updated either to be smaller, i.e, allowing more samples to be accepted, or to be larger, i.e., resulting in more rejections (see Lines 11 and 13).

Algorithm 9: Adaptive β and $\mathfrak{n}$ update

input :

- Number of different regimes N_K
- Regularization factor $\mathfrak{r}$
- Length of the Markov chain $N_{\text{chain}}^{\text{RWM}}$
- Initial values for the noise factor $\mathfrak{n}$ and inverse temperature β

output:

- Global optimizer Γ^*

```
1  Choose or generate an initial Γ^[0], β^[0] and n^[0].
2  for r = 1 : N_chain^RWM do
3      Propose new sample Γ'
4      Accept/Reject-procedure
5      if accept then
6          N_accept = N_accept + 1
7          if L^r(Γ^[r−1]) > L^r(Γ') then
8              N_accept+lowerEnergy = N_accept+lowerEnergy + 1
9      if mod(r, 1000) = 0 then
10         if N_accept < 90 then
11             n = n · 0.85
12         else if N_accept > 140 then
13             n = n · 1.05
14         else
15             n = n
16         N_accept = 0
17         N_accept+lowerEnergy = 0
18     if mod(r, 1000) = 500 then
19         if N_accept − N_accept+lowerEnergy ≥ N_accept · 0.25 then
20             β = β · 1.111;
21         N_accept = 0
22         N_accept+lowerEnergy = 0
23 Return Γ^(N_chain^RWM)
```

Note that the regarded number of iteration steps, considered before the dynamical tuning procedure, is reduced from 1000 to 500. This is due to the additional update of the parameter β (see Lines 18-20), which also takes place after 1000 chain members have been proposed.

A simultaneous adaptive update might cause immense changes in the percentage of accepted samples, potentially leading to a loss of control. Thus, the tuning of each variable is executed at different lengths of the Markov chain. In detail, this means that the first update of β is set to take place after 500 iteration steps and then regularly executed after 1000 iteration steps, resulting in shifted, by 500 steps, update procedures. As discussed in Subsection 3.5.1, the inverse temperature β needs to be increased very slowly in order to circulate in the whole sample space.

This can be achieved by keeping the temperature value relatively big at the beginning of the sampling process. Nevertheless, it is important to steadily increase β, causing the accepted samples in general to have lower energy values than the current chain member and, thus, to converge to a Boltzmann distributed sample, minimizing the functional $\mathbf{L}^{\mathrm{r}}(\Gamma, \Theta)$. In particular, β is updated if less than 75% of the accepted samples have lower energy, i.e.,

$$\frac{N_{\text{accept}} - N_{\text{accept+lowerEnergy}}}{N_{\text{accept}}} \geq 0.25. \tag{B.1}$$

This adaptive simulated annealing scheme has been numerically investigated on synthetic as well as real observations for a non-stationary k-means model (see model example defined in (3.2)) [30]. The results are very promising and suggest that the proposed Metropolis optimization ansatz is a good alternative to the FEM clustering approach [52].

An implementation, written in C++, of the subspace optimization of the model parameter Γ for a purely time-dependent data set for f^{kmeans} (see definition in (3.2)), with the introduced RWM algorithm including the simulated annealing scheme, can be found on `http://www.dewiljes.de/dewiljes/Jana.html`. The source code is commented and a corresponding example of synthetic data sets can also be downloaded. The reader is referred to Subsection 3.5.1 of this thesis, and [30] for further discussion on the theory of the methodology and the numerical approach. Note that the source code can easily be generalized to suit other direct model functions f by replacing the deterministic expression (used for the energy computations in e.cpp) of a minimal Θ^* for fixed Γ with the $\Theta^{[s]}$, determined via the subspace optimization.

C

NUMERICAL RESULTS

Extended information on the numerical results corresponding to the synthetical examples of Chapter 4 and the arctic sea ice application of Chapter 5 are presented.

C.1 TOY EXAMPLE 1

In order to estimate out-of-sample state probabilities it is necessary to select model parameters that optimally describe the affiliations Γ^*. The mAICc values computed for different model choices for the first toy example (see Chapter 4) are given in Table 2. The underlying model is assumed to be stationary and homogenous, i.e., $N_K = 1$. Note that N_C does not have to be set for the stationary, homogenous case. The definition of $\bar{u}^{\text{syn}}(t, j)$ can be found in Section 4.1.

$\mathbf{mAICc}(1, -, N_M^\Gamma, \bar{u}^{\text{syn}}(t,j), f^\Gamma)$	f	N_M^Γ
-35233	Markov (see (3.22))	1
-29285	Markov (see (3.22))	0
-33156	logit (see (3.22))	0

Table 2: *The mAICc values attained for different models describing the dynamics of Γ^* associated with the process underlying the artificial data in Section 4.1 are given.*

C.2 TOY EXAMPLE 2

An optimal model used to describe the affiliation process Γ^* (details in Section 4.2) is selected via mAICc. The resulting values are displayed in the following table.

$\mathbf{mAICc}(1, -, N_M^\Gamma, \bar{u}^{\text{syn}}(t,j), f^\Gamma)$	f	N_M^Γ
-33814	Markov (see (3.22))	1
-32644	logit (see (3.22))	1
-6684.2	Markov (see (3.22))	0
-6484.3	logit (see (3.22))	0

Table 3: *The mAICc values attained for different models describing the dynamics of Γ^* associated with the process underlying the artificial data in Section 4.2 are given.*

C.3 ARCTIC SEA ICE APPLICATION

The matrices P_e^{k*} with $e \in \{1, 2, N_E\}$ and $k \in \{1, 2, N_K^*\}$ corresponding to the optimal model $\mathcal{M}^{\text{Markov}}(3, 70, 0, [\text{neigh}_{\text{ice}}, \text{temp}, \text{CO}_2]^\top)$ are

$$\begin{aligned} P_1^{1*} &= \begin{bmatrix} 0.1795 & -0.1795 \\ 0.1795 & -0.1795 \end{bmatrix}, P_1^{2*} = \begin{bmatrix} 0.2002 & -0.2002 \\ 0.2002 & -0.2002 \end{bmatrix}, \\ P_1^{3*} &= \begin{bmatrix} 0.0130 & -0.0130 \\ 0.0130 & -0.0130 \end{bmatrix}. \end{aligned} \tag{C.1}$$

$$\begin{aligned} P_2^{1*} &= \begin{bmatrix} -0.1485 \cdot 10^{-16} & 0.9704 \cdot 10^{-16} \\ -0.3188 \cdot 10^{-16} & -0.1442 \cdot 10^{-16} \end{bmatrix}, \\ P_2^{2*} &= \begin{bmatrix} -0.0887 & 0.0887 \\ -0.0887 & 0.0887 \end{bmatrix}, P_2^{3*} = \begin{bmatrix} -0.0152 & 0.0152 \\ -0.0152 & 0.0152 \end{bmatrix}. \end{aligned} \tag{C.2}$$

$$\begin{aligned} P_3^{1*} &= \begin{bmatrix} -0.1301 \cdot 10^{-16} & -0.0726 \cdot 10^{-16} \\ -0.0889 \cdot 10^{-16} & -0.4077 \cdot 10^{-16} \end{bmatrix}, \\ P_3^{2*} &= \begin{bmatrix} -0.0032 & 0.0032 \\ -0.0032 & 0.0032 \end{bmatrix}, P_3^{3*} = \begin{bmatrix} -0.0051 & 0.0051 \\ -0.0051 & 0.0051 \end{bmatrix}. \end{aligned} \tag{C.3}$$

Note that matrices P_0^{1*}, P_0^{2*} and P_0^{3*} have already been given in Subsection 5.3.2 (see (5.8)).The entries of the matrices (see (C.1), (C.2) and (C.3)) are used to determine the statistical impact of the associated external factors $u_e(t, j)$ (for details see Subsection 5.3.3).

The following tables displays all the resulting mAIC values of all 126 executed runs, the corresponding information such as the choice of used explicit external factors $u_e(t, j)$ and the resulting optimal number of regimes N_K^* as

well as N_C^*) the maximal number of allowed transitions in the considered time interval are given.

mAICc	N_K^*	N_C^*	$u(t,j)$
−569680	3	70	[neigh$_{ice}$, temp, CO_2]
−568450	3	80	[neigh$_{ice}$, neigh$_{land}$, CO_2, NAO, AO]
−567270	3	80	[neigh$_{ice}$, neigh$_{land}$, CO_2]
−566570	3	80	[neigh$_{ice}$, neigh$_{land}$, CO_2, NAO]
−566340	3	80	[neigh$_{ice}$,CO_2,NAO,AO]
−565980	3	80	[neigh$_{ice}$,temp]
−565660	3	80	[neigh$_{ice}$,CO_2,AO]
−564470	3	80	[neigh$_{ice}$,CO_2,NAO]
−564420	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, NAO, AO]
−564220	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, NAO]
−563340	3	80	[neigh$_{ice}$,temp,CO_2, NAO]
−563130	3	80	[neigh$_{ice}$,temp,CO_2, NAO, AO]
−563030	3	80	[neigh$_{ice}$, neigh$_{land}$, CO_2, AO]
−563000	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, AO]
−562940	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2]
−562870	3	80	[neigh$_{ice}$, NAO, AO]
−561550	3	80	[neigh$_{ice}$]
−561350	3	80	[neigh$_{ice}$, temp, CO_2, AO]
−559430	3	80	[neigh$_{ice}$, temp, NAO]
−559180	3	70	[neigh$_{ice}$, neigh$_{land}$]
−559130	3	80	[neigh$_{ice}$, neigh$_{land}$, AO]
−558820	3	70	[neigh$_{ice}$, CO_2]
−558780	3	80	[neigh$_{ice}$, neigh$_{land}$, NAO]
−558530	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, NAO, AO]
−558460	3	80	[neigh$_{ice}$, neigh$_{land}$, temp]
−558080	3	80	[neigh$_{ice}$, neigh$_{land}$, NAO, AO]
−558030	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, NAO]
−557700	3	80	[neigh$_{ice}$, neigh$_{land}$, temp, AO]
−555950	3	80	[neigh$_{ice}$, NAO]
−553810	3	80	[neigh$_{ice}$, AO]
−552720	3	80	[neigh$_{ice}$, temp, AO]
−549940	3	80	[neigh$_{ice}$, temp, NAO, AO]
−546450	3	80	[neigh$_{land}$, CO_2]
−545860	3	80	[neigh$_{land}$, CO_2, NAO]
−545620	3	80	[CO_2, NAO]

−544470	3	80	[CO_2]
−543700	3	80	[CO_2, NAO, AO]
−541660	3	80	[$\text{neigh}_{\text{land}}$, CO_2, NAO, AO]
−540400	3	80	[temp, CO_2, NAO]
−539140	3	80	[temp, CO_2, NAO, AO]
−539130	3	80	[$\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO]
−537790	3	80	[$\text{neigh}_{\text{land}}$, CO_2, AO]
−537410	3	80	[temp, CO_2, AO]
−537250	3	80	[CO_2, AO]
−536650	3	80	[temp, CO_2]
−536330	3	80	[$\text{neigh}_{\text{land}}$, temp, CO_2, NAO]
−533940	3	80	[$\text{neigh}_{\text{land}}$, temp, CO_2]
−533480	3	80	[$\text{neigh}_{\text{land}}$, temp, CO_2, AO]
−531880	3	80	[AO]
−531770	3	80	[$\text{neigh}_{\text{land}}$]
−531230	3	80	[NAO, AO]
−529860	3	80	[$\text{neigh}_{\text{land}}$, NAO]
−529670	3	80	[$\text{neigh}_{\text{land}}$, AO]
−529260	3	80	[temp]
−529220	3	80	[$\text{neigh}_{\text{land}}$, temp]
−529070	3	80	[temp, AO]
−526380	3	80	[NAO]
−524760	3	70	[temp, NAO]
−523880	3	70	[temp, NAO, AO]
−505190	3	80	[$\text{neigh}_{\text{land}}$, temp, NAO]
−498930	3	80	[$\text{neigh}_{\text{land}}$, temp, NAO, AO]
−451810	3	80	[$\text{neigh}_{\text{land}}$, NAO, AO]
−448800	3	60	[$\text{neigh}_{\text{land}}$, temp, AO]

Table 4: *The* ***mAICc***$(N_K^*, N_C^*, 0, u(t,j), f^{Markov})$ *values of models* $\mathcal{M}^f(N_K^*, N_C^*, 0, u(t,j))$ *for different external factor combinations of entries* $u_e(t,j) \in \mathcal{E}$ *are displayed.*

mAICc	N_K^*	N_C^*	$u(t,j)$
−533610	2	80	[$\text{neigh}_{\text{land}}$,$CO_2$, NAO]
−533550	2	80	[$\text{neigh}_{\text{land}}$,AO]
−533410	2	80	[$\text{neigh}_{\text{land}}$, temp, CO_2, NAO]
−533090	2	80	[$\text{neigh}_{\text{land}}$, temp, CO_2]
−533090	2	80	[$\text{neigh}_{\text{land}}$,$CO_2$]
−532630	2	70	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, CO_2, NAO]

−532480	2	70	[neigh$_{ice}$, neigh$_{land}$, CO_2,AO]
−532330	2	80	[neigh$_{land}$,CO_2, AO]
−532230	2	80	[neigh$_{ice}$, neigh$_{land}$, CO_2]
−532080	2	70	[neigh$_{land}$, temp, CO_2, NAO, AO]
−532070	2	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, NAO, AO]
−532040	2	70	[neigh$_{land}$, temp, CO_2, AO]
−532040	2	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2]
−532010	2	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, NAO]
−531550	2	70	[neigh$_{ice}$, neigh$_{land}$, CO_2, NAO, AO]
−531320	2	80	[neigh$_{ice}$, neigh$_{land}$, temp, CO_2, AO]
−531320	2	80	[CO_2, NAO]
−531320	2	80	[CO_2]
−531090	2	70	[temp, CO_2]
−531020	2	80	[neigh$_{ice}$, CO_2, NAO]
−530930	2	80	[CO_2, NAO,AO]
−530890	2	80	[CO_2,AO]
−530880	2	70	[neigh$_{land}$, CO_2, NAO, AO]
−530850	2	80	[neigh$_{ice}$, temp, CO_2, AO]
−530830	2	80	[neigh$_{ice}$, temp, CO_2, NAO, AO]
−530780	2	80	[neigh$_{ice}$, [neigh$_{land}$]
−530680	2	80	[neigh$_{ice}$,CO_2, AO]
−530630	2	80	[temp, CO_2, NAO]
−530420	2	70	[temp, CO_2, NAO, AO]
−530390	2	80	[neigh$_{ice}$, CO_2]
−530280	2	80	[neigh$_{ice}$, temp, CO_2, NAO]
−529800	2	80	[neigh$_{ice}$,CO_2, NAO, AO]
−529750	2	80	[neigh$_{ice}$, temp, CO_2]
−529710	2	70	[neigh$_{ice}$, neigh$_{land}$, AO]
−529540	2	80	[temp, CO_2, AO]
−528930	2	80	[temp, NAO]
−528810	2	80	[neigh$_{ice}$, neigh$_{land}$, temp, NAO]
−528070	3	80	[neigh$_{land}$]
−527450	2	80	[temp]
−527410	2	80	[temp, AO]
−526110	2	80	[AO]
−525960	2	80	[neigh$_{ice}$, AO]
−525640	2	80	[neigh$_{ice}$, temp, NAO]
−525060	2	80	[neigh$_{ice}$, temp, AO]
−522890	2	80	[neigh$_{ice}$, temp]

−521020	2	30	[NAO, AO]
−519780	2	60	[$\text{neigh}_{\text{ice}}$,$\text{neigh}_{\text{land}}$, temp, AO]
−514100	2	50	[$\text{neigh}_{\text{ice}}$]
−513390	2	30	[$\text{neigh}_{\text{ice}}$,$\text{neigh}_{\text{land}}$, temp, NAO, AO]
−511720	2	40	[$\text{neigh}_{\text{ice}}$,$\text{neigh}_{\text{land}}$, temp]
−508370	2	40	[$\text{neigh}_{\text{ice}}$, NAO]
−501700	2	30	[$\text{neigh}_{\text{ice}}$,$\text{neigh}_{\text{land}}$, NAO]
−501480	2	20	[NAO]
−497870	3	80	[$\text{neigh}_{\text{land}}$, temp, AO]
−497260	3	80	[$\text{neigh}_{\text{land}}$, temp, NAO]
−493410	2	30	[$\text{neigh}_{\text{land}}$, temp]
−491340	2	80	[temp, NAO, AO]
−491230	2	80	[$\text{neigh}_{\text{land}}$, temp, NAO, AO]
−483250	4	5	[$\text{neigh}_{\text{ice}}$, temp, NAO, AO]
−482530	2	80	[$\text{neigh}_{\text{land}}$, NAO]
−471320	2	20	[$\text{neigh}_{\text{ice}}$, NAO, AO]
−463620	3	50	[$\text{neigh}_{\text{land}}$, NAO, AO]
−456820	2	80	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, NAO, AO]

Table 5: *The* $\boldsymbol{mAICc}(N_K^*, N_C^*, 1, u(t,j), f^{Markov})$ *values of models* $\mathcal{M}^f(N_K^*, N_C^*, 1, u(t,j))$ *for different external factor combinations of entries* $u_e(t,j) \in \mathcal{E}$ *are displayed.*

mAICc	f^Γ	N_M^Γ	$u^\Gamma(t,j)$
−158560	Markov	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_3]
614.34	Markov	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_3]
−140410	logit	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_3]
−111114	logit	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_3]
−140260	Markov	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_1]
−29285	Markov	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_1]
−135860	logit	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_1]
−163629	logit	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO, period_1]
−152066	Markov	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO]
7105.2	Markov	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO]
−117990	logit	1	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO]
−65443	logit	0	[$\text{neigh}_{\text{ice}}$, $\text{neigh}_{\text{land}}$, temp, CO_2, NAO, AO]

Table 6: *The* $\boldsymbol{mAICc}(1, -, N_M^\Gamma, u^\Gamma(t,j), f^\Gamma)$ *values attained for different models describing the dynamics of* Γ^**, where* $N_K^\Gamma = 1$*, are displayed. Note that a logistic model with memory has* $\gamma_1(t-1,j)$ *and* $\gamma_3(t-1,j)$ *as additional external factors.*

NOTATION

The notation index is organized as follows: the numbers and sizes are listed separately as their notation is very similar. The remaining notations are listed in order of appearance in the thesis. To improve readability, the titles of corresponding chapters or sections are indicated. Moreover, the abbreviations used in the thesis are listed at the end of the notation index.

Numbers and sizes

N_S	number of states s_i (associated index: i, page 11)
N_L	number of data samples (associated index: i, page 11)
N_E	number of explicit external factors (associated index: e, page 11)
$N^{\text{SVM}}_{\text{boxconstraint}}$	regularization parameter corresponding to an SVM run (page 13)
$N^{\text{ANN}}_{\text{neurons}}$	number of employed neurons for an ANN run (page 18)
N_T	length of observed time series $\pi(t, j)$ for fixed location j (associated index: t, page 21)
N_{ens}	(associated index: l, page 21)
N_J	space dimension of observations $\pi(t, j)$ for all time steps t (associated index: j, page 21)
$N_{s_i}(t, j)$	number of cells j currently (at time t) in state s_i (page 22)
N_F	number external factors (associated index: e, page 23)
N_I	number of implicit external factors (associated index: e, page 23)
N_M	memory depth (page 28)
N_K	number of local stationary homogenous models θ_k (associated index: k, page 34)
N_C	maximal number of allowed transitions of the affiliation processes $\gamma_k(t, j)$ for fixed j (page 38)
$N^{\text{FEM}}_{\text{anneal}}$	number of annealing steps used for the FEM framework (page 41)
$N^{\text{FEM}}_{\tau\text{ol}}$	(page 41)
$N^{\text{FEM}}_{\text{basis}}$	number of finite elements used for the discretization (page 43)
$N^{\text{RWM}}_{\text{chain}}$	length of generated Markov chain (page 46)
N_{ϕ_k}	degree of a polynomial of ϕ_k (page 53)
N_{pred}	prediction depth (page 57)
N^{Γ}_K	number of regimes for characterization of Γ^* (page 57)
N^{Γ}_S	number of model states for characterization of Γ^* (page 57)
$N_{T_{\text{train}}}$	time-wise length of training data (page 59)

$N_{\mathrm{K}}^{\mathrm{syn}}$	artificially chosen number of local stationary homogenous models θ_k^{syn} (page 60)
$N_{\mathrm{C}}^{\mathrm{syn}}$	artificially chosen maximal number of transitions of the synthetic affiliation processes $\gamma_k^{\mathrm{syn}}(t,j)$ (page 60)
N_{dummy}	auxiliary quantity of Algorithm 5 (page 62)
$N_{\mathrm{anneal}}^{\mathrm{ANN}}$	number of annealing steps used for an ANN run (page 65)
N_{C}^{Γ}	maximal number of transitions for characterization of Γ^* (page 91)
N_M^{Γ}	memory depth of model for characterization of Γ^* (page 91)
N_Y	number of years (page 97)
N_{accept}	number of accepted samples (page 115)
$N_{\mathrm{accept+lowerEnergy}}$	number of accepted samples with lower energy (page 116)
N_{K}^*	optimal (with respect to the mAICc) maximal number of local stationary models (page 120)
N_{C}^*	optimal (with respect to the mAICc) maximal number of transitions (page 121)

Support vector machines

s_i	discrete state (page 11)
$\sigma(l)$	discrete process (page 11)
$\mathbb{N}$	positive integers (page 11)
$\mathbb{R}$	real numbers (page 11)
$u(l)$	data samples (page 11)
$y(l)$	class assignments (page 11)
$\langle \cdot, \cdot \rangle$	dot product (page 12)
w	vector describing the hyperplane (page 12)
m	hyperplane variable (page 12)
$\|\| \cdot \|\|_2$	Euclidean norm (page 12)
$D(\lambda)$	dual problem (page 13)
$\lambda(l)$	Lagrange multipliers (page 13)
$\varsigma(l)$	slack variable (page 13)
$\Upsilon(\cdot)$	projection function affiliated with kernel function (page 14)
$\mathcal{K}(\cdot, \cdot)$	kernel function (page 14)
$\mathcal{K}^{\mathrm{poly}}(\cdot, \cdot)$	polynomial kernel function (page 14)
$\mathcal{K}^{\mathrm{RBF}}(\cdot, \cdot)$	radial basis kernel function (page 14)
$\mathcal{K}^{\mathrm{MLP}}(\cdot, \cdot)$	multilayer perceptron kernel function (page 15)

Artificial neural networks

$\mathcal{W}$	vector of weights corresponding to a neuron (page 15)
$\mathfrak{b}$	bias of a neuron (page 15)
$\Psi(\cdot)$	transfer function (page 16)
$\Psi^{\tanh}(t)$	hyperbolic tangent transfer function (page 16)
$\Psi^{\text{sigmoid}}(t)$	logistic transfer function (page 16)
$\Psi^{\text{rectifier}}(t)$	rectifier transfer function (page 16)
$\mathcal{N}(N^{\text{ANN}}_{\text{neurons}})$	MLP with one hidden layer with $N^{\text{ANN}}_{\text{neurons}}$ (page 18)

Logit models

$C_i[u(t,j), B^i]$	utility measure (page 18)
B^i	logistic model parameter (page 18)
b^i_e	eth entry of vector B^i (page 18)
$\xi^i(t,j)$	error process of utility measure (page 19)
$\mathbb{P}[\cdot]$	probability function (page 19)

Discrete spatio-temporal dynamical process

$\sigma(t,j,l)$	spatio-temporal dynamical process (page 21)

Ensemble data and external factors

$\omega(j,l)$	microscopic cell (page 21)
$\tilde{\pi}_i(t,j)$	empirical state probability (page 22)
$\delta_{s_i}(\cdot)$	Kronecker delta for the value s_i (page 22)
$\pi_i(t,j)$	state probability in location $\omega(j,l)$ at time t (page 22)
$\pi(t,j)$	vector of states probabilities (page 23)
$\bar{u}(t,j)$	all influencing external factors (page 23)

Implicit external factors

$u(t,j)$	vector of explicit external factors (page 23)
$u_e(t,j)$	explicit external factor (page 23)
$\mathcal{U}$	space of explicit external factors $u(t,l)$ (page 23)
$u^{\text{unres}}(t,j)$	vector of implicit external factors (page 23)
$u^{\text{unres}}_e(t,j)$	implicit external factor (page 23)

Model distance function

$\theta(\bar{u}(t,j))$	unknown model parameter dependent on $\bar{u}(t,j)$ (page 28)

Ω	parameter space containing $\theta(\bar{u}(t,j))$ (page 28)
$f(\cdot)$	denotes a general direct mathematical model (page 28)
f^{kmeans}	kmeans direct mathematical model function (page 28)
$\epsilon(t,j)$	error term of simple model example (page 28)
f^{logit}	logistic direct mathematical model function (page 28)
$\theta^{\text{logit}}(B(t,j),u(t,j))$	logistic direct mathematical model parameter (page 28)
$\zeta(t,j)$	error term of logistic model distance function (page 28)
$B(t,j)$	non-stationary non-homogenous logistic parameter (page 28)
$g(\cdot)$	model distance function (page 29)
$d(\cdot,\cdot)$	metric (page 29)
$\mathbf{E}(\cdot)$	expected value (page 29)
$\mathbf{L}(\cdot)$	averaged clustering functional (page 29)

Markov model

$P(\bar{u}(t,j))$	transition matrix dependent on all external factors (page 30)
$\varepsilon(t,j)$	error term associated with transition matrix $P(\bar{u}(t,j))$ (page 31)
$P_0(t,j)$	matrix used in the linear combination equal to $P(\bar{u}(t,j))$ (page 31)
$P_e(t,j)$	matrix used in the linear combination equal to $P(\bar{u}(t,j))$ (page 31)
$\mu(t,j)$	vector of expected values of vector $\bar{u}(t,l)$ (page 31)
$R_\alpha(\bar{u}(t,j))$	Taylor expansion error component (page 31)
α	multi-index (page 31)
$P(t,j,u(t,j))$	approximation of $P(\bar{u}(t,j))$ (page 32)
f^{Markov}	Markov direct mathematical model function (page 32)

Interpolation

$\theta(t,j,\bar{u}(t,j))$	non-stationary, non-homogenous model parameter (page 33)
$\theta^{\text{kmeans}*}(t,j)$	optimal model parameter for model function f^{kmeans} (page 33)
$\Theta(u(t,j))$	vector of stationary homogenous model parameters (page 34)
$\theta_k(u(t,j))$	stationary homogenous model parameter (page 34)
$\Gamma(t,j)$	matrix of affiliations $\gamma_k(t,j)$ (page 34)
$\gamma_k(t,j)$	affiliations corresponding to local model $\theta_k(u(t,j))$ (page 34)
$\mathbf{L}(\cdot,\cdot)$	interpolated version of $\mathbf{L}(\cdot)$ (page 34)
$\mathbf{L}_j(\cdot,\cdot)$	one summand for a fixed location j of interpolated average clustering functional (page 34)
B_k	local logit model parameter (page 35)
B_k^i	vector of stationary and homogenous logit model parameter matrix B_k (page 35)
$P^k(u(t,j))$	local Markov model parameter matrix (page 35)

$P_0^k, \ldots, P_{N_E}^k$	matrices used in the linear combination equal to $P^k(u(t,l))$ (page 35)
$P(u(t,j))$	vector of model matrices $P^k(u(t,j))$ (page 35)
$\mathbf{1}$	auxiliary column vector contain entries equal to one (page 35)
$\mathbf{0}$	auxiliary column vector contain entries equal to zero (page 35)
$\left\{P_e^k\right\}_{n,m}$	entry of matrix P_e^k in nth row and mth column (page 36)

Spatial and temporal persistence

$\Gamma^*(t,j)$	global optimizer with respect to $\mathbf{L}(\Gamma(t,j), \Theta(u(t,j)))$ (page 37)
$\gamma_k^*(t,j)$	global optimizer with respect to $\mathbf{L}(\Gamma(t,j), \Theta(u(t,j)))$ for fixed Θ (page 37)
$W^{1,2}([1,N_T])$	Sobelev space (page 37)
$L^2([1,N_T])$	Lebesgue space (page 37)
$\mathbf{L}^{\mathfrak{r}}(\Theta,\Gamma)$	Tikhonov-regularized averaged clustering functional (page 37)
$\mathfrak{r}$	regularization factor (page 37)
$\mathbf{L}_j^{\mathfrak{r}}(\Theta,\Gamma)$	one summand for a fixed location j of Tikhonov-regularized averaged clustering functional (page 38)
$\left\Vert \frac{\partial \gamma_k}{\partial t} \right\Vert^2_{L^2([1,N_T])}$	L^2-norm (page 38)
$\vert \cdot \vert_{BV(1,N_T)}$	bounded variation (BV) half-norm (page 38)

Spatial relations

$\mathrm{neigh}(j)$	direct neighbor cells of cell j (page 40)

Numerical approach and computational complexity

$\Theta^*(u(t,j))$	global optimizer with respect to $\mathbf{L}(\Gamma(t,j), \Theta(u(t,j)))$ (page 40)
$\mathcal{M}^f(N_K, N_C, N_M, u(t,j))$	model (page 41)
$P^{[s]}(u(t,j))$	current approximation of optimal $P^*(u(t,j))$ (page 41)
$\Gamma^{[s]}$	current approximation of optimal Γ^* process (page 41)
$\mathbf{L}_{\min}$	auxiliary variable of Algorithm 1 (page 42)
$\mathcal{F}_{\mathbf{L}^{\mathfrak{r}},\beta}(\Gamma)$	Boltzmann distribution of regularized averaged clustering functional $\mathbf{L}^{\mathfrak{r}}(\Theta,\Gamma)$ with fixed Θ (page 45)
$Z(\mathbf{L}_j^{\mathfrak{r}})$	normalizing constant of Boltzman distribution (page 45)
β	inverse temperature variable of Boltzmann distribution (page 45)
$q(\cdot,\cdot)$	proposal density (page 46)
$\mathfrak{a}(\Gamma^{[r-1]}, \Gamma')$	acceptance rate (page 46)
Γ'	newly proposed sample of Γ^* (page 47)
η	noise used to generate new sample (page 47)

$\psi_k(t,j)$	auxiliary processes for fixed k (page 48)
$\psi(t,j)$	vector of processes $\psi_k(t,j)$ (page 48)
$\beta^{[r]}$	currently used inverse temperature parameter (page 50)
$\mathfrak{n}$	noise factor (page 52)
$\mathcal{O}(\cdot)$	Big O notation (page 52)

Model selection

$\mathcal{L}(N_K, N_C, N_M, u(t,j))$	loglikelihood (page 53)
$\phi_k\left(\cdot,\ldots,\cdot\middle\vert N_{\phi_k}\right)$	parametric (conditional) probability density function (page 53)
$\mathbf{mAIC}(\cdot,\cdot,\cdot,\cdot,\cdot)$	modified version of AIC (page 54)
$\vert\mathcal{M}^f(N_K, N_C, N_M, u(t,j))\vert$	number of involved parameters of a model (page 54)
$\vert\mathcal{M}^{\text{Markov}}(N_K, N_C, N_M, u(t,j))\vert$	number of involved parameters for Markov model (page 54)
$\vert\Gamma\vert$	number of free variables required to reconstruct Γ (page 54)
$\vert\mathcal{M}^{\text{logit}}(N_K, N_C, N_M, u(t,j))\vert$	number of involved parameters for a logistic model (page 54)
$\mathbf{mAICc}(\cdot,\cdot,\cdot,\cdot,\cdot)$	corrected version of modified AIC (page 56)

Self-containing predictive models

$\hat{\pi}(t,j)$	prediction of observation $\pi(t,j)$ (page 56)
$\hat{\Gamma}(t,j)$	prediction of future affiliations (page 57)
$P_0^\Gamma, \ldots, P_{N_E}^\Gamma$	model matrices for characterization of Γ^* (page 57)
B^Γ	logistic model parameter for characterization of Γ^* (page 58)

Test model systems

$\sigma^{\text{syn}}(t,j,l)$	synthetic dynamical process (page 59)
$\pi^{\text{Markov}}(t,j)$	approximation of $\pi^{\text{syn}}(t,j)$ computed via Markov model (page 59)
$u^{\text{syn}}(t,j)$	synthetic explicit external factors (page 59)
$\Gamma^{\text{syn}}(t,j)$	synthetic affiliation process (page 60)
$P^{\text{syn}}(t,j,u(t,j))$	synthetic transition matrix (page 60)
$P^{k\ \text{syn}}(u(t,j))$	synthetic model parameter matrix associated with affiliation $\gamma_k^{\text{syn}}(t,j)$ (page 60)
$\pi^{\text{syn}}(t,j)$	synthetic data (page 61)
$P_{N_E+1}^{k\ \text{syn}}, \ldots, P_{N_E+N_I}^{k\ \text{syn}}$	synthetic model matrices corresponding to implicit external factors (page 61)
$\bar{u}^{\text{syn}}(t,j)$	vector of all synthetic external factors (page 61)
$\gamma_k^{\text{syn}}(t,j)$	synthetic affiliation associated with θ_k^{syn} (page 62)
dummy0	auxiliary vector of Algorithm 5 containing only zeros (page 62)

dummy1	auxiliary vector of Algorithm 5 containing only ones (page 62)
$P_0^{k\ \mathrm{syn}}, \ldots, P_{N_E}^{k\ \mathrm{syn}}$	synthetic model matrices corresponding to explicit internal factors (page 63)
$\pi_i^{\mathrm{syn}}(t,j)$	ith vector entry of synthetic data $\pi^{\mathrm{syn}}(t,j)$ (page 63)
$\mathcal{N}(N_{\mathrm{neurons}}^{\mathrm{ANN}})$	network with $N_{\mathrm{neurons}}^{\mathrm{ANN}}$ neurons (page 65)
$\pi^{\mathcal{N}(N_{\mathrm{neurons}}^{\mathrm{ANN}})}(t,j)$	approximation of $\pi^{\mathrm{syn}}(t,j)$ computed via $\mathcal{N}(N_{\mathrm{neurons}}^{\mathrm{ANN}})$ (page 65)
$\varpi_{\mathrm{rel}}(\tau)$	relative mean prediction error (page 68)
$\varpi(j,\tau)$	prediction error term dependent (page 68)

External factors

$\mathcal{E}$	set of considered explicit external factors (page 79)
$\mathrm{neigh}_{\mathrm{ice}}(t,j)$	averaged state probability of all neighbors of cell j (page 80)
$\mathrm{neigh}_{\mathrm{land}}(t,j)$	land percentages surrounding cell j (page 80)

Parameter identification and results

$\mathcal{M}^{f^\Gamma,\Gamma}(N_K^\Gamma, N_C^\Gamma, N_M^\Gamma, u^\Gamma(t,j))$	model for Γ^* (page 91)
$\mathcal{E}^\Gamma$	set of explicit external factors considered for parametrization of Γ^* (page 91)
$\mathcal{I}(e,k)$	statistical impact (page 95)
$\mathcal{I}_{\mathrm{rel}}(e,k)$	relative statistical impact (page 95)

Non-stationary non-homgenous Markov regression

$\mathbf{p}$	entries of unknown model matrices in a collumn vector (page 107)
$\mathbf{G}$	quadratic part of quadratic programming problem (page 107)
$\mathbf{g0}$	linear part of quadratic programming problem (page 107)
$\mathbf{vec}(P_e)$	entires of model matrix P_e in a vector (page 108)
$g0(t,j)$	summand of vector $\mathbf{g0}$ (page 109)
$\mathbf{diag}(\cdot)$	matrix containing on smaller matrices along the diagonal (page 109)
$G(t,j)$	summand of matrix $\mathbf{G}$ (page 109)
$\mathbf{p}^k$	entries of unknown model matrices in a column vector for regime k (page 110)
$\mathbf{G}^k$	quadratic part of quadratic programming problem for regime k (page 110)
$\mathbf{g0}^k$	linear part of quadratic programming problem for regime k (page 110)
$\mathbf{CE}(N_M)$	matrix corresponding to equality constraint (page 111)
$\mathbf{ce0}$	vector corresponding to equality constraint (page 111)

$\mathcal{Q}(\mathbf{Id}_{N_S})$	auxiliary matrix (page 111)
CI	matrix corresponding to inequality constraint (page 112)
ci0	vector corresponding to inequality constraint (page 112)

Abbreviations

SDEs	Stochastic Differential Equation
MCMC	Markov Chain Monte Carlo
SVMs	Support Vector Machines
RBF	Radial Basis Function
MLPs	Multilayer Perceptrons
ANNs	Artificial Neural Networks
AIC	Akaike Information Criterion
mAIC	modified Akaike Information Criterion
mAICc	corrected modified Akaike Information Criterion
PDEs	Partial Differential Equations
FEM	Finite Elemente Methode
IIA	Independence of Irrelevant Alternatives
i.i.d.	independent and identically distributed

BIBLIOGRAPHY

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] H. Akaike. Information Theory and an Extension of the Maximum Likelihood Principle. In B. Petrov and F. Caski, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado Budapest, 1973.

[3] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[4] M. H. P. Ambaum, B. J. Hoskins, and D. B. Stephenson. Arctic Oscillation or North Atlantic Oscillation? *Journal of Climate*, 14(16):3495–3507, 2001.

[5] S. E. Amstrup, E. T. DeWeaver, D. C. Douglas, B. G. Marcot, G. M. Durner, C. M. Bitz, and D. A. Bailey. Greenhouse gas mitigation can reduce sea ice loss and increase polar bear persistence. *Nature*, 468(7326):955–958, 2010.

[6] A. G. Barnston and R. E. Livezey. Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns. *Monthly Weather Review*, 115(6):1083–1126, 1987.

[7] M. H. Beale, M. T. Hagan, and H. B. Demuth. *Neural Network Toolbox^{TM}: User's Guide*. The MathWorks, Inc., 2014.

[8] T. Bengtsson and J. E. Cavanaugh. An improved Akaike information criterion for state-space model selection. *Computational Statistics and Data Analysis*, 50(10):2635–2654, 2006.

[9] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[10] F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):637–654, 1973.

[11] C. Blume, K. Matthes, and I. Horenko. Supervised Learning Approaches to Classify Stratospheric Warming Events. *Journal of the Atmospheric Sciences*, 69(6):1824–1840, 2012.

[12] A. Bordes, S. Ertekin, J. Weston, and L. Bottou. Fast Kernel Classifiers with Online and Active Learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

[13] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[14] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. John Wiley and Sons, 2008.

[15] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 2nd edition, 2001.

[16] M. J. Brodzik and K. W. Knowles. EASE-Grid: A Versatile Set of Equal-Area Projections and Grids. In M. Goodchild and A. J. Kimerling, editors, *Discrete Global Grids*. Santa Barbara, California USA: National Center for Geographic Information and Analysis, 2002.

[17] K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference: A practical Information-Theoretic Approach*. Springer; 2nd edition, 2002.

[18] E. N. Cassano, J. J. Cassano, M. E. Higgins, and M. C. Serreze. Atmospheric impacts of an Arctic sea ice minimum as seen in the Community Atmosphere Model. *International Journal of Climatology*, 34(3):766–779, 2014.

[19] J. E. Cavanaugh, S. L. Davies, and A. A. Neath. Discrepancy-Based Model Selection Criteria Using Cross Validation. In F. Vonta, M. Nikulin, N. Limnios, and C. Huber, editors, *Statistical Models and Methods for Biomedical and Technical Systems*, Statistics for Industry and Technology, chapter 33, pages 473–485. Birkhauser, Boston, Massachusetts, 2008.

[20] N. I. Center. National Ice Center Arctic sea ice charts and climatologies in gridded format. *Edited and compiled by F. Fetterer and C. Fowler. Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.*, 2006 (updated 2009).

[21] D. Chen and X. Yuan. A Markov Model for Seasonal Forecast of Antarctic Sea Ice. *Journal of Climate*, 17(16):3156–3168, 2004.

[22] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49(4):327–335, 1995.

[23] J. Chipman. The foundations of utility. *Econometrica*, 28(2):193–224, 1960.

[24] R. Collobert and S. Bengio. Links between perceptrons, MLPs and SVMs. In *Proceedings of the 21st International Conference on Machine Learning*, page 23. ACM, 2004.

[25] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[26] N. Cressie and C. K. Wikle. *Statistics for Spatio-Temporal Data*. Wiley, Hoboken, NJ, 2011.

[27] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, 2000.

[28] G. B. Dantzig and M. N. Thapa. *Linear Programming 1 : Introduction*. Springer Series in Operations Research. Springer, 1997.

[29] G. B. Dantzig and M. N. Thapa. *Linear Programming 2: Theory and Extensions*. Springer Series in Operations Research. Springer, 2003.

[30] J. de Wiljes, A. Majda, and I. Horenko. An Adaptive Markov Chain Monte Carlo Approach to Time Series Clustering of Processes with Regime Transition Behavior. *Multiscale Modeling and Simulation*, 11(2):415–441, 2013.

[31] J. de Wiljes, L. Putzig, and I. Horenko. Discrete nonhomogeneous and nonstationary logistic and Markov regression models for spatiotemporal data with unresolved external influences. *Communications in Applied Mathematics and Computational Science*, 9(1):1–46, 2014.

[32] P. Del Moral, A. Doucet, and A. Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society*, 68(3):411–436, 2006.

[33] A. Dobson and A. Barnett. *Introduction to Generalized Linear Models*. Chapman and Hall/CRC., 2008.

[34] A. Einstein. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Annalen der Physik*, 322(8):549–560, 1905.

[35] I. Eisenman, T. Schneider, D. S. Battisti, and C. M. Bitz. Consistent Changes in the Sea Ice Seasonal Cycle in Response to Global Warming. *Journal of Climate*, 24(20):5325–5335, 2011.

[36] M. H. England, S. McGregor, P. Spence, G. A. Meehl, A. Timmermann, C. Wenju, A. S. Gupta, M. J. McPhaden, A. Purich, and A. Santoso. Recent intensification of wind-driven circulation in the Pacific and the ongoing warming hiatus. *Nature Climate Change*, 4(3):222–227, 2014.

[37] J. Falck. Adopting a Bayesian framework to multidimensional cluster modeling. Master's thesis, Free University Berlin, 2010.

[38] R. B. Fuller. *Synergetics*. Macmillan New York, 1975.

[39] N. Ganesan, K. Venkatesh, M. A. Rama, and A. M. Palani. Application of neural networks in diagnosing cancer disease using demographic data. *International Journal of Computer Applications*, 1(26):76–85, 2010.

[40] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[41] J. Gill. *Generalized Linear Models*. SAGE Publications, Inc., 2001.

[42] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33, 1983.

[43] M. F. Goodchild. Geographical grid models for environmental monitoring and analysis across the globe (panel session). In *Proceedings of GIS/US '94 Conference, Phoenix, Arizona*, 1994.

[44] R. W. Gray. Exact Transformation Equations for Fuller's World Map. *Cartographica*, 32(3):17–25, 1995.

[45] J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[46] M. T. Hagan and M. B. Menhaj. Training feedforward networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994.

[47] O. Häggström. *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press, 2002.

[48] W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[49] R. P. Hauser and D. Booth. Predicting Bankruptcy with Robust Logistic Regression. *Journal of Data Science*, 9(4):565–584, 2011.

[50] R. Heikes and D. A. Randall. Numerical Integration of the Shallow-Water Equations on a Twisted Icosahedral Drid. Part I: Basic Design and Results of Tests. *Monthly Weather Review*, 123(6):1862–1880, 1995.

[51] R. Heikes and D. A. Randall. Numerical Integration of the Shallow-Water Equations on a Twisted Icosahedral Grid. Part II: A Detailed Description of the Grid and an Analysis of Numerical Accuracy. *Monthly Weather Review*, 123(6):1881–1887, 1995.

[52] I. Horenko. Finite Element Approach to Clustering of Multidimensional Time Series. *SIAM Journal on Scientific Computing*, 32(1):62–83, 2010.

[53] I. Horenko. Nonstationarity in Multifactor Models of Discrete Jump Processes, Memory, and Application to Cloud Modeling. *Journal of the Atmospheric Sciences*, 68(7):1493–1506, 2011.

[54] I. Horenko. On Analysis of Nonstationary Categorical Data Time Series: Dynamical Dimension Reduction, Model Selection and Applications To Computational Sociology. *Multiscale Modeling and Simulation*, 9(4):1700–1726, 2011.

[55] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[56] C.-W. Hsu and C.-J. Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[57] A. Hu, C. Rooth, R. Bleck, and C. Deser. NAO influence on sea ice extent in the Eurasian coastal region. *Geophysical Research Letters*, 29(22), 2002.

[58] IPCC. *Climate Change 2007: The Physical Science Basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: Cambridge, NY., 2007.

[59] J. Jacod and A. Shiryaev. *Limit Theorems for Stochastic Processes*. Springer, 2nd edition, 2002.

[60] M. A. Katsoulakis, A. J. Majda, and A. Sopasakis. Multiscale Couplings in Prototype Hybrid Deterministic/Stochastic Systems: Part I, Deterministic Closures. *Communications in Mathematical Sciences*, 2(2):255–294, 2004.

[61] M. A. Katsoulakis, A. J. Majda, and A. Sopasakis. Multiscale Couplings in Prototype Hybrid Deterministic/Stochastic Systems: Part II, Stochastic Closures. *Communications in Mathematical Sciences*, 3(3):453–478, 2005.

[62] M. A. Katsoulakis, A. J. Majda, and A. Sopasakis. Hybrid deterministic stochastic systems with microscopic look-ahead dynamics. *Communications in Mathematical Sciences*, 8(2):409–437, 2010.

[63] J. E. Kay, M. M. Holland, and A. Jahn. Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters*, 38(15), 2011.

[64] A. J. Kimerling, K. Sahr, D. White, and L. Song. Comparing Geometrical Properties of Global Grids. *Cartography and Geographic Information Science*, 26(4):271–288, 1999.

[65] R. Kindermann and J. L. Snell. *Markov Random Fields and Their Applications*. American Mathematical Society, 1980.

[66] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by Simulated Annealing. *Science*, 220(4598):671–680, 1983.

[67] Y. Kosaka and S.-P. Xie. Recent global-warming hiatus tied to equatorial Pacific surface cooling. *Nature*, 501(7467):403–407, 2013.

[68] V. Kurkova. Kolmogorov's theorem and multilayer neural networks. *Neural Networks*, 5(3):501–506, 1992.

[69] J. Lawrence. *Introduction to Neural Networks*. California Scientific Software Press, 1994.

[70] P. Lesage, F. Glangeaud, and J. Mars. Applications of autoregressive models and time–frequency analysis to the study of volcanic tremor and long-period events. *Journal of Volcanology and Geothermal Research*, 114(3-4):391–417, 2002.

[71] W. D. Li and C. A. McMahon. A simulated annealing-based optimization approach for integrated process planning and scheduling. *International Journal of Computer Integrated Manufacturing*, 20(1):80–95, 2007.

[72] J. Liu and C. Sabbatti. Simulated sintering: Markov chain Monte Carlo with spaces of varying dimensions. *Bayesian Statistics*, 6:386–413, 1998.

[73] J. S. Liu. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

[74] R. D. Luce. *Individual choice behavior: a theoretical analysis*. Wiley, New York, 1959.

[75] C. Manski and D. McFadden. *Structural Analysis of Discrete Data and Econometric Applications*. Cambridge: MIT Press., 1981.

[76] C. Marchetti, P. S. Meyer, and J. H. Ausubel. Human population dynamics revisited with the logistic model: How much can be modeled and predicted? *Technological Forecasting and Social Change*, 52(1):1–30, 1996.

[77] E. Marinari and G. Parisi. Simulated Tempering: A New Monte Carlo Scheme. *Europhyics Letters*, 19(6):451–458, 1992.

[78] W. Maslowski, J. C. Kinney, M. Higgins, and A. Roberts. The Future of Arctic Sea Ice. *Annual Review of Earth Planetary Science*, 40:625–654, 2012.

[79] The MathWorks, Inc. *Statistical Toolbox™ User's Guide*.

[80] P. McCullagh and J. A. Nelder. *Generalized linear models. Monographs on Statistics and Applied Probability.* Chapman and Hall/CRC, 1989.

[81] D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 105–142, 1974.

[82] R. C. Merton. Theory of rational option pricing. *The Bell Journal of Economics and Management Science*, 4(1):141–183, 1973.

[83] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[84] P. Metzner, L. Putzig, and I. Horenko. Analysis of persistent nonstationary time series and applications. *Communications in Applied Mathematics and Computational Science*, 7(2):175–229, 2012.

[85] T. Ogawa, H. Sonoda, A. Ishiwa, and Y. Shigeta. An Application of Autoregressive Model to Pattern Discrimination of Brain Electrical Activity Mapping. *Brain Topography*, 6(1):3–11, 1993.

[86] T. J. O'Kane, R. J. Matear, M. A. Chamberlain, J. S. Risbey, B. M. Sloyan, and I. Horenko. Decadal variability in an OGCM Southern Ocean: Intrinsic modes, forced modes and metastable states. *Ocean Modelling*, 69:1–21, 2013.

[87] T. J. O'Kane, J. S. Risbey, C. Franzke, I. Horenko, and D. P. Monselesan. Changes in the Metastability of the Midlatitude Southern Hemisphere Circulation and the Utility of Nonstationary Cluster Analysis and Split-Flow Blocking Indices as Diagnostic Tools. *Journal of the Atmospheric Sciences*, 70(3):824–842, 2013.

[88] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition, 2014.

[89] J. E. Overland and M. Wang. When will the summer Arctic be nearly sea ice free. *Geophysical Research Letters*, 40(10):2097–2101, 2013.

[90] N. S. Pillai, A. M. Stuart, and A. H. Thiéry. Optimal scaling and diffusion limits for the Langevin algorithm in high dimensions. *The Annals of Applied Probability*, 22(6):2320–2356, 2012.

[91] I. G. Rigor, J. M. Wallace, and R. L. Colony. Response of Sea Ice to the Arctic Oscillation. *Journal of Climate*, 15(18):2648–2663, 2002.

[92] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.

[93] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

[94] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, 16(4):351–367, November 2001.

[95] L. Roux, D. Racoceanu, N. Loménie, M. Kulikova, H. Irshad, J. Klossa, F. Capron, C. Genestie, G. L. Naour, and M. N. Gurcan. Mitosis detection in breast cancer histological images An ICPR 2012 contest. *Journal of Pathology Informatics*, 4(8), 2013.

[96] R. Sadourny, A. Arakawa, and Y. Mintz. Integration of the nondivergent barotropic vorticity equation with an icosahedral-hexagonal grid for the sphere. *Monthly Weather Review*, 96(6):351–356, 1968.

[97] K. Sahr. Location coding on icosahedral aperture 3 hexagon discrete global grids. *Computers, Environment and Urban Systems*, 32(3):174–187, 2008.

[98] K. Sahr, D. White, and A. J. Kimerling. Geodesic Discrete Global Grid Systems. *Cartography and Geographic Information Science*, 30(2):121–134, 2003.

[99] A. Schweiger, R. Lindsay, J. Zhang, M. Steele, H. Stern, and R. Kwok. Uncertainty in modeled Arctic sea ice volume. *Journal of Geophysical Research*, 116(C8), 2011.

[100] J. A. Screen, I. Simmonds, C. Deser, and R. Tomas. The Atmospheric Response to Three Decades of Observed Arctic Sea Ice Loss. *Journal of Climate*, 26(4):1230–1248, 2013.

[101] S. Shalev-Shwartz and N. Srebro. SVM optimization: Inverse dependence on training set size. *Proceedings of the 25th international conference on Machine learning*, pages 928–935, 2008.

[102] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, 2004.

[103] M. E. Silva, S. Barbosa, L. Antunes, and C. Rocha. Quantile regression analysis of Arctic sea ice extent. In *EGU General Assembly Conference Abstracts*, volume 11, page 4210. Wien, Austria, 2009.

[104] J. S. Singarayer, J. L. Bamber, and P. J. Valdes. Twenty-First-Century Climate Impacts from a Declining Arctic Sea Ice Cover. *Journal of Climate*, 19(7):1109–1125, 2006.

[105] J. P. Snyder. An Equal-Area Map Projection For Polyhedral Globes. *Cartographica*, 29(1):10–21, 1992.

[106] E. H. Spanier. *Algebraic Topology*. Springer, 1st edition, 1981.

[107] R. H. Stewart. *Introduction to Physical Oceanography*. Orange Grove Texts Plus, 2009.

[108] J. C. Stroeve, V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. Meier. Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters*, 39(16), 2012.

[109] D. W. Stroock. *An Introduction to Markov Processes*. Springer, 2005.

[110] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, 2005.

[111] R. Tareghian and P. Rasmussen. Analysis of Arctic and Antarctic sea ice extent using quantile regression. *International Journal of Climatology*, 33(5):1079–1086, 2013.

[112] J. Thuburn. A PV-Based Shallow-Water Model on a Hexagonal-Icosahedral Grid. *Monthly Weather Review*, 125(9):2328–2347, 1997.

[113] K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[114] I. B. Vapnyarskii. Lagrange multipliers. In M. Hazewinkel, editor, *Encyclopedia of Mathematics*. Springer, Heidelberg, 2001.

[115] S. A. Vavasis. Quadratic programming is in NP. *Information Processing Letters*, 36(2):73–77, 1990.

[116] J. von Neumann. Various techniques used in connection with random digits. *National Bureau of Standards Applied Math Series*, 11:36–38, 1951.

[117] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[118] M. Wang and J. E. Overland. A sea ice free summer Arctic within 30 years: An update from CMIP5 models. *Geophysical Research Letters*, 39(18), 2012.

[119] D. White, A. J. Kimerling, and W. S. Overton. Cartographic and Geometric Components of a Global Sampling Design for Environmental Monitoring. *Cartography and Geographic Information Systems*, 19(1):5–22, 1992.

[120] D. White, A. J. Kimerling, K. Sahr, and L. Song. Comparing area and shape distortion on polyhedral-based recursive partitions of the sphere. *International Journal of Geographical Information Science*, 12(8):805–827, 1998.

[121] F. E. Wickman, E. Elvers, and K. Edvarson. A system of domains for global sampling problems. *Geografiska Annaler*, 56(3/4):201–212, 1974.

[122] P. Wolfe. The simplex method for quadratic programming. *Econometrica*, 27(3):382–398, 1959.

[123] L. Ying, D. Xu, and Z.-P. Liang. On Tikhonov regularization for image reconstruction in parallel MRI. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 1056–1059, 2004.

DECLARATION

I hereby declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or any other institute of tertiary education. Further, I have, to the best of my knowledge, acknowledged all sources in the text and a list of references is given.

Berlin, June 2014

Jana de Wiljes