Università
Ca'Foscari
Venezia

Master's Degree programme
in Environmental Sciences

Final Thesis

—

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Forecasting water temperature in Northern Adriatic lagoons: a functional data approach

Supervisor
Ch. Prof. Carlo Gaetan

Supervisor
Ch. Prof. Roberto Pastres

Graduand
Daniel Glaser
Matriculation number
877142

Academic Year
2020/2021

Abstract

The management of extensive aquacultures requires easily applicable modelling solutions for the prediction of various water quality variables, especially given the fast-changing climatic conditions due to an emerging climate crisis. Functional data analysis offers an approach for data driven models, which are able to produce reliable forecasts. Various publications have outlined this methodology remarking its huge potential and broad field of possible application. While it is common practice in some fields of medicine, as of now it finds little reference and scarce examples of implementation in environmental literature.

This work explores the applicability of Functional Auto-Regressive models (FAR) for forecasting the hourly evolution of relevant water quality variables. As a proof-of-concept this methodology was used to forecast half-hourly water temperature fluctuations in shallow water transition systems located in the Northern Adriatic, Italy, namely the Lagoon of Venice and the Marinetta Lagoon in the Po delta. In order to achieve this goal a two step modelling approach was developed: 1) daily mean values were forecasted using ARIMAX model, 2) daily oscillations around the means were predicted using a FARX model. Air temperature and salinity were added as external predictors. Findings show that modelling results for one-day forecasts are of high predictive power. Better 2-4 days ahead prediction could be obtained by improving the trend estimation.

These results indicate that FARX models are a sound and flexible class of data driven models, which could be used as a tool for the prediction of highly dynamic water quality variables in shallow coastal lagoons, such as water temperature. Further work, however, is needed to test FARX on other relevant variables, i.e. dissolved oxygen and salinity and to extend the forecast window. Weekly forecast of these variables would, indeed, be very useful to support the management of culture based fishery and aquaculture in these ecosystems, as they could provide early warning concerning adverse events, e.g. heat waves and ipoxias.

Contents

List of abbreviations

AIC	Akaike information criterion
AR	Autoregression
ARIMAX	Autoregressive integrated moving average with external variable
ARMA	Autoregression moving average
ARMAX	Autoregression moving average with extrenal variable
ARPAV	Agenzia Regionale per la Prevenzione e Protezione Abientale del Veneto
BIC	Bayesian information criterion
CNR-ISMAR	National Research Council, Marine Sciences Institute
EFPC	Estimated functional principal component
FAR	Functional autoregression
FARX	Functional autoregression with external variabel
FDA	Functional data analysis
FPC	Functional principle component
FTU	Formazin turbidity unit
MA	Moving average
MLE	Maximum likelihood estimation
PC	Principle component
PSU	Practical Salinity Unit
REML	Restricted maximum likelihood
SSR	Sum of squared residuals

1 Introduction

The global population is rapidly increasing, having doubled in the past 50 years to now approximately 7.7 billion (Roser et al., 2013). This trend is expected to continue until the end of the century, peaking at 11 billion global citizens (UN Department for Economic and Social Affairs, 2019). Simultaneously a substantial decrease of farmland quality can be noticed in many regions of the world (Jie et al., 2002), which is why efficient and sustainable food production is becoming increasingly important in order to meet the Sustainable Development Goals (UN General Assembly, 2015).

Unsurprisingly, the so-called Blue Economy, a term which denominates the economic exploitation of the marine environment, is currently experiencing an increase in popularity amongst scientists and entrepreneurs (Smith-Godfrey, 2016). Large-scale food production in coastal and off-shore areas represents one of the most promising ways to tackle food security (Hussain et al., 2018; Fredheim and Langan, 2009). To this regard, the European Union has launched the long term strategic initiative Blue Growth, which includes the development of aquaculture as one of its pillars (European Commission. Directorate-General for Maritime Affairs and Fisheries, 2012). Due to high pressure by both overfishing and the effects of climate change wild fish stocks are decreasing in many places over the world (Hilborn et al., 2020; Capuzzo et al., 2018; Edgar et al., 2018; Vasilakopoulos et al., 2014; Brander, 2007) Simultaneously aquacultures have experienced a rapid upsurge in numbers (Tacon, 2020).

For centuries coastal areas and transition ecosystems, such as coastal lagoons, have been regarded as highly important sources of fish, shellfish and game. In particular, different forms of cultured based fishery and extensive aquaculture have been practised since the 15th century in enclosed portions of the Lagoon of Venice and of other Northern Adriatic lagoons, named ”Valli da pesca” . In the last decades, shellfish farming and, more specifically, the farming of the allochthonous clam *Ruditapes philippinarum* has become the most relevant halieutic resource. However, this farming activity is facing some challenges, due to the increasing climate variability related to climate change (Ghezzo et al., 2018). Therefore, it is important to both understand the environmental conditions necessary for successful farming, as well as being able to predict the development of water quality variables, such as water temperature, salinity and dissolved oxygen concentration, which could cause sub-lethal and lethal stress to the farmed species. As regards the latter, the simultaneous occurrence of heat waves and hypoxic conditions could cause mass mortalities: predicting these events could be crucial for shellfish and fish farmers.

This is a challenging task, as the dynamic of these variables is driven by complex physical, chemical and ecological processes, such as tidal exchanges, heat exchanges with the atmosphere, wind driven mixing and resuspension, sediment oxygen consumption or oxygen depletion due to algal blooms. As a result, most quality variables show an underlying daily pattern, which, however, is often masked by a high level of ”noise”. These patterns can be simulated using both process based and data driven models: the former, however, are computationally expensive and require highly skilled personnel to be developed and operated. Therefore, in a management context, data driven models could be a better option for producing real time or near real time site-specific predictions.

The focus of this work will be placed upon the latter aspect, developing an innovative data driven model approach, based on functional data analysis, for 1-2 day ahead forecasting of water quality variables in a lagoon, based on processing of real time data.
This approach could be used for implementing precision shellfish farming within lagoons such as the Lagoon of Venice by providing early warning concerning variables which can cause stress and mass mortality. These predictions could be used for mitigating the consequences of these extreme events, by harvesting or, if possible moving the stock, to less affected areas.

The advantage of functional auto-regression is that it can use continuous functions as regression objects. Consequentially the modelling resolution can be adjusted at will. This is especially useful for virtually continuously observed variables as is common for environmental parameters.

The following chapter 2 will revise existing literature on different approaches for modelling waterbodies. Chapter 3 lays out the methodology and mathematical concepts regarding functional auto-regressive (FARX) models. As a proof of concept, FARX models were applied to forecast water temperature, whose dynamics is affected mainly by physical processes, namely, heat exchanges with the atmosphere and heat transport due to tidal mixing. This was undertaken in two different lagoons in the Northern Adriatic each for various seasons (see chapter 4). The results are presented in chapter 5, chapter 6 discusses the findings and chapter 7 provides suggestions for further research and possible applications of this method.

2 Water models

This chapter will illustrate the current state of the art of water models focussing on water temperature models. Section 2.1 will give a short introduction to the classification of mathematical models. Subsequently, water temperature models applied to waterbodies such as rivers and estuaries are reviewed (section 2.2). Section 2.3 gives an account of existing water models for the Northern Adriatic.

2.1 Classification of models

There are several types of models that can be applied to describe a water body (Benyahya et al., 2007): first of all it has to be distinguished whether a model is process based or data driven. The former simulates the dynamics of a system by describing the relationships among state variables using sets of ODE (Ordinary Differential Equations) or PDE (Partial Differential Equations), taking into account the external forcings, i.e. the exchange of matter and energy with the surroundings. These models require a detailed knowledge about a system and a high amount of data concerning input parameters and external forcings, such as topography, hydrology, weather forcings and others. Process based biogeochemical and ecological models are routinely applied, for example, in oceanography (e.g. Lopes et al. (2018) and Dube and Jayaraman (2008)).
Their main disadvantages, which constrain their applicability, are the large computational effort as well as the need of highly skilled personnel to develop and maintain them.

Data driven models, on the other hand, analyse observed data and identify reoccurring patterns. They can be further divided into parametric and non-parametric models depending on whether the amount of parameters is finite or not. Non-parametric models do not have a previously defined structure or amount of parameters, but develop based on the training data instead (e.g. artificial neural networks). This flexible approach is becoming increasingly popular as computational power advances and facilitates their application. A downside to this approach is the lack of information about how the model works making an interpretation difficult.

A final distinction for parametric models can be made into regressive and dynamic. Regression models express one variable as a function of one or more independent variables. The relation between input and output variable can be linear (linear regression) or of a more general nature (non-linear regression). Other noteworthy types of parametric models are autoregressive and periodic autoregressive ones which can be labelled as dynamic models. In contrast to regressive models they regard a variables autocorrelation and compute the development of a variable based on its own past values. Other external variables can be added. If a periodic component is existent then periodic autoregression should be used, which splits the data into a long-term (seasonal) component and short-term impulses (errors).

2.2 Water temperature models

Given the abundance of water on earth's surface and its relevance for life, it is not surprising that there is a great number of publications dealing with various aspects of it. A large part of that research is dedicated to modelling its dynamic in different water bodies, such as local streams, subterraneous aquifers or seas. Since not every model is suitable for every type of application, there is no one superior model but, rather, a wide variety of models suited for different applications.

Coastal aquatic systems are often highly interconnected with their surroundings and involve many relevant processes. Water temperature can be simulated using an energy budget approach, making use of the laws of thermodynamics to derive the models equations (Benyahya et al., 2007). Such a model was also applied to the Lagoon of Venice (Dejak et al., 1992).
A deterministic, finite-element approach was applied by Umgiesser and his colleagues (Umgiesser (1997), Umgiesser et al. (2004), Ferrarin and Umgiesser (2005), Umgiesser et al. (2014), Maicu et al. (2018)). In the latter paper, a 3D finite element model is used for simulating the hydrodynamic transport and the heat budget, producing as a result a numerically solved computer simulation. Ferrarin and Umgiesser (2005) applied the same approach to a lagoon in Sardinia from which they conclude that the water temperature in shallow lagoons can be efficiently modelled by using the air temperature as a proxy for the radiative forcing.

Chen et al. (1998) showed that this is also a major forcing for other water bodies such as streams. They implemented detailed information about stream topography and characteristics of the riparian foliage to construct a model for simulating the hourly water temperature based on shading and air temperature. Vaz et al. (2005), on the other hand introduced a 2-dimensional water temperature model based on the hydrodynamic water transport equations using the water temperature as a tracer for river inflow. However, their results showed a discrepancy of $3°C$ between observed and computed water temperature.

Regressive models for water temperatures also frequently use the correlation between air and water temperature to deduct one from the other. According to the literature, linear regression is suitable for predicting the seasonal evolution of water temperatures. Seasonal autoregressive models are suitable for data with strong seasonality. Caissie et al. (1998) for example model the water temperature in a small Canadian stream by approximating the seasonal component with a Fourier series and then use various autoregressive models to describe the residuals while adding air temperatures as an external predictor.
Mohseni et al. (1998) expand this idea using four parameters to better account for weekly minima and maxima, making it a non-linear regressive model which is able to describe the typical levelling off of water temperatures at high and low temperatures more precisely. They explain this phenomenon with the effects of freezing and evaporative cooling respectively.

This work is intended to explore the application of functional data analysis (FDA) to the modelling of the daily temperature pattern in a transition water body. Based on the results presented in Ullah and Finch (2013), 84 articles on FDA were published between 1995 and 2010, More than half of these papers (54%) concerned applications in biomedicine, biomechanics, medicine, psychology and neurology and environmental sciences, in the broader sense, (biology, ecology, meteorology, environmental studies and agriculture) accounted for only 20%.

This study supports the findings from an own provisional literature research, that this method is barely mentioned across the corresponding literature. Several scientific search engines and data bases (GoogleScholar, Scopus, Springer and Web of Science) were searched using the keywords ”FARX model”, ”functional autoregression”, ”functional data analysis” and ”functional data application” for publications in environmental sciences which use functional data analysis for modelling.
One of the rare examples is the study of Mestekemper et al. (2010) who use a functional data approach to forecast the hourly development of the water temperature of the river Wupper in Germany. The method applied by Mestekemper et al. is to a large extend analogous to the one presented in this work apart from the varying ecosystem that is to be examined.

2.3 Water models for the Northern Adriatic

Due to its special role as a city which exists in symbiosis with the surrounding lagoon, Venice has early on evoked a scientific interest in the latter. Consequentially models of the Lagoon of Venice date back as far as the 1970s (Di Silvio and D’Alpaos (1972), Chignoli and Rabagliati (1973), Sguazzero et al. (1978)). While these models are still unrefined from a modern point of view, many others have developed more sophisticated models over the years which exploit the great knowledge of the lagoon that has been gained by now.

In 1997 Umgiesser developed a finite element model for the lagoon of Venice which implicitly accounts for the bottom topography. He also demonstrates that the overall water circulation is driven southward by the north-eastern Bora wind system. In 2004 the model is updated by Umgiesser et al. to include temporal evolution of the variables. The new version is apt to represent dispersion processes which affect salinity and water temperature accurately.
Canu et al. extend the finite element model in 2003 to be able to analyse the ecosystem’s response to changes in the physical conditions. The previous model is upgraded by an energy budget model to incorporate water temperature changes and several ecological models which account for nutrients, organic matter and dissolved oxygen, all of which are important factors in an estuarine ecosystem.

In 1998 Bergamasco et al. combine a model for the Adriatic Sea and one for the Venetian lagoon, arguing that both systems are coupled and mutually influence each other. In their study they use two hydrodynamic models of different resolution (lower for the Adriatic basin), which incorporate tides, winds and other fluxes as forcings. A different study by Ferrarin et al. (2017) shows the extensive influence of the three major lagoon systems in the northern Adriatic (Marano, Venice and the Po-Delta) on circulation and sediment transport within the whole basin. These studies are some of the few that couple numerical models for lagoons and the Adriatic Sea even though some of the lagoons have a water flushing time (ratio between water exchange and volume) of one day, hence should be regarded as strongly interacting with the Adriatic (Umgiesser et al., 2014).

Umgiesser et al. (2014) introduce the SHYFEM model, a 3-dimensional extension of the finite element model which includes vertical z-layers. It has been developed by the CNR-ISMAR (National Research Council, Marine Sciences Institute) in Venice and was successfully applied to several lagoons in Europe (Ferrarin and Umgiesser (2005), Umgiesser et al. (2014)).
Another recent study (Maicu et al., 2018) uses the earlier mentioned SHYFEM model to tackle the complex Po-Delta system which is defined by several river branches and lagoons. Noteworthy is the great interconnectivity between the different water bodies which demands great knowledge about local conditions for calibrating a hydrodynamical model.

3 Methodology

This chapter provides a basic understanding of the applied methodology and its underlying mathematical concepts. This is not a purely mathematical thesis but rather focusses on the development and practical application of a functional model. Therefore no claim is made that the methods below are defined perfectly unambiguous and described in a way to endure a mathematicians scrutiny.

3.1 Auto-Regressive Moving Average models

An assumption that can be made for many environmental parameters is that the future values of a variable highly depend on its current value. Therefore it is only sensible to incorporate past values into the calculation of future values. This principle is used for so-called autoregressive models such as the widely used AR(p) model.

An autoregressive process y at the time t of the order p can be described as follows:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \varepsilon_t \tag{3.1.1}$$

Here α denotes a constant and ϕ_i the modelling parameters for the autoregressive model. The error (labelled as ε_t) is assumed to be a white noise, i.e. $E(\varepsilon_t) = 0$, $var(\varepsilon_t) = \sigma^2$ and ε_t is an independent random variable. The order p indicates how many past values are taken into consideration for the model calculation.

In order to preserve stationarity of the process some constraints apply to the modelling parameters. For an AR(1) process $|\phi_1| < 1$ has to be valid. The constraints for an AR(2) process are $|\phi_1| < 1, \;\; \phi_1 + \phi_2 < 1, \;\; \phi_2 - \phi_1 < 1$. The general formulation for an AR(p) process is more complex and can be found in the corresponding literature. Since modelling software like R usually regards these constraints automatically, this should be of little concern (Shumway and Stoffer, 2000).

If the examined variable does not only depend on its past values but also on another (independent) variable, this can be expanded to an ARX(p,l) model by including the current value and the last l lags of an external variable x:

$$y_t = \alpha + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{k=0}^{l} \beta_k x_{t-k} + \varepsilon_t \tag{3.1.2}$$

The coefficients β_k are the modelling parameters for the external input x.

If the forecasting errors also seem to be relevant for a process then an autoregressive moving average model (ARMA) can be applied. Additionally to the already described AR(p) process, it includes a moving average term (MA(q)) which regards the q last forecasting errors:

$$y_t = \alpha \; + \sum_{i=1}^{p} \phi_i y_{t-i} \; + \sum_{j=1}^{q} \eta_j \varepsilon_{t-j} \; + \; \varepsilon_t \tag{3.1.3}$$

If it's necessary to make a time series stationary, then it can be differentiated once or more often before modelling. The reversion of this process is here referred to as integration.
The equation for differentiating y_t once ($d = 1$) is as follows:

$$y'_t = y_t - y_{t-1} \tag{3.1.4}$$

An ARMA model with a d-times differentiated time series gets the additional notation I(d), making it an ARIMA(p,d,q) model.
To extend the described ARIMA(p,d,q) model by an external predictor x with l lags (cf. equation 3.1.2 for an ARX process) the corresponding term has to be added:

$$y_t = \alpha \ + \sum_{i=1}^{p} \phi_i y_{t-i} \ + \sum_{j=1}^{q} \eta_j \varepsilon_{t-j} \ + \ \sum_{k=0}^{l} \beta_k x_{t-k} \ + \ \varepsilon_t \tag{3.1.5}$$

Like before β_k denotes the modelling parameters for the external variable. In the following all the above mentioned variants will be referred to as ARIMAX model, since that is the most inclusive term.

3.2 Functional Data Analysis

Like mentioned in section 3.1, autoregression and moving average can be useful tools to predict environmental variables, but in some cases an ARMAX model is not suitable for data sets with a high sampling frequency. Such a model regards all observed values as discrete values that need to be forecasted individually. Besides causing unnecessary high computation effort it has also proven to be inefficient in forecasting for example hourly values for multiple days (Mestekemper et al., 2010).

Kokoszka and Reimherr (2017) give an updated account of a new concept of data examination. The field of functional data analysis deals with variables that can be described by smooth curves and performs a statistical analysis on a set of such curves.
When applied to a time series, the data set is split into sets of equal time length (e.g.: days, weeks or years). N denotes the total number of subsets while T stands for the total amount of discrete observations points in each subset (for simplicity's sake we assume evenly spaced out data). Figure 3.1 shows a time series with high sampling frequency (once per minute over 7 days) sectioned into seven subsets ($N = 7$) of one day length each ($T = 1440\ min$).

Every subset $Y_n = \{Y_n(t), t = 1, \cdots, T\}$ can be approximated by a basis-expansion:

$$Y_n(t) \approx \sum_{m=1}^{M} c_{nm} \Psi_m(t) \tag{3.2.1}$$

$\Psi_m(t)$ is a standard set of basis functions like the Fourier basis or splines and c_{nm} are the M corresponding coefficients (see figure 3.2). This approximation creates a smooth function which goes through all observation points and replaces

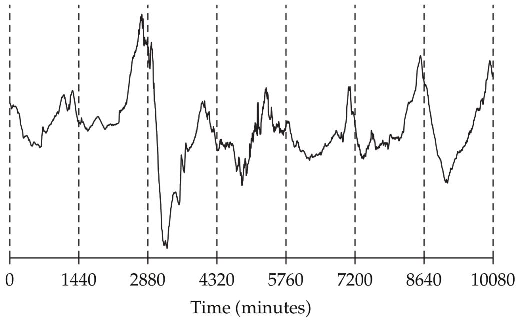


Figure 3.1: "The horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT." Adapted from Kokoszka (2012).

the discrete data Y_n. It also serves to replace Y_n by a smaller M-dimensional vector of coefficients c_{nm}.

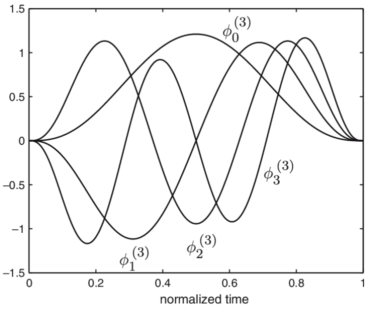


Figure 3.2: Example of the first four basis functions of the Fourier expansion. Adapted from Biess et al. (2006).

The expansion displayed in equation 3.2.1, commonly uses deterministic basis functions. Instead, to gain a maximum compression, one can find the optimal basis functions specially for the given data. This is done via functional principal component analysis:

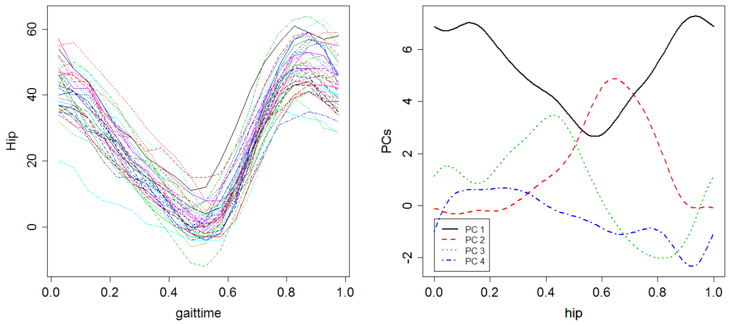
$$Y_n - \overline{Y}_N(t) \approx \sum_{m=1}^{M} \xi_{nm} \upsilon_m(t) \tag{3.2.2}$$

The centred function $Y_n - \overline{Y}_N(t)$ is approximated by M functions υ_m, the so called estimated functional principal components (EFPCs). The coefficients ξ_{nm} are called the score of Y_n to the respective basis function υ_m. The score quantifies how much the respective basis function contributes to the shape of Y_n. It is obtained as fitting coefficients when matching the set of EFPCs to the centred function.

The EFPCs are defined to be a set of orthonormal trigonometric functions, meaning that:

$$\int \upsilon_m(t)\upsilon_i(t)\ dt = \begin{cases} 0 & \text{if} \quad m \neq i, \\ 1 & \text{if} \quad m = i. \end{cases} \tag{3.2.3}$$

The first EFPC (υ_1) summarizes the main variability around the mean function, while the following EFPCs summarize the main remaining variability orthogonal to the previous ones. The m-th EFPC (υ_m) represents the dominant remaining variation orthogonal to all the previous EFPCs ($\upsilon_1, \upsilon_2, \cdots, \upsilon_{m-1}$). The total variability can be written as a sum of variabilities explained by each basis function. Thus it is possible to quantify how much one EFPC contributes to explain the overall variability. In practice a threshold of 95% or 99% resemblance is often chosen to exclude further EFPCs which only describe a fraction of the variability.
This selection process sometimes results in M being a much smaller value than when using a standard set of basis expansions (cf. 3.2.1), which again results in much less variables to handle while loosing no or only little information. It is important to note that for both approximation methods the observations do not necessarily need to be evenly spaced out. Figure 3.3 shows a typical application for functional data analysis: 39 observations of the same process performed by different probands (3.3a) are approximated by using the first four EFPCs (3.3b).



(a) Hip Angles observed over gait cycle for 39 children.

(b) First four EFPCs approximating the hip gait data.

Figure 3.3: Adopted from Cao (2019).

Through this functional approximation the previously T observations of every subset are now reduced to M parameters which, combined together, describe the curve approximating the development of the variable within that interval as one functional object. Thus reduced, the approximated curves for one or more interval can be used for further analysis such as modelling or predicting future values.

3.3 Functional Auto-Regressive models

The in section 3.2 introduced functional approach can be exploited to expand an AR model, making it a Functional Auto-Regressive model (FAR). As mentioned in Kokoszka and Reimherr (2017) a FAR(1) model for the n^{th} subset can be written as:

$$Y_n = \Phi(Y_{n-1}) + \varepsilon_n \tag{3.3.1}$$

This is the functional equivalent to equation 3.1.1 with Y_n being the approximated function (as mentioned in 3.2) which represents the values of the input and output function. ε_n is a sequence of iid mean zero elements of L^2. The integral operator Φ transforms a function into another function and resembles the modelling parameter ϕ_i from equation 3.1.1. It is defined as follows:

$$\Phi(Y_n(t)) = \int Y_n(s)\varphi(t,s)\ ds \tag{3.3.2}$$

Equation 3.3.1 can thus be rewritten as:

$$Y_n(t) = \int Y_{n-1}(s)\varphi(t,s)\ ds + \varepsilon_n(t)$$

For the sake of a more simple notation this representation of a FAR model assumes the mean function to be zero. Otherwise the mean function μ has to be deducted from the observation. It can be estimated by $\hat{\mu} = \bar{Y}_N = N^{-1}\sum_{n=1}^{N} Y_n$.

In order to account for external forcings that are not contained in Y_n the FAR model can be extended by one or more additional external variables X_n to become a FARX model. The functional equivalent to equation 3.1.2 is denoted as follows:

$$Y_n = \Phi(Y_{n-1}) + B(X_n) + \varepsilon_n \tag{3.3.3}$$

The added term X_n is another functional approximation this time representing the external variable on day n. As before Φ, B is another integral operator. The full representation of the FARX model can also be written as:

$$Y_n(t) = \int Y_{n-1}(s)\varphi(t,s)\ ds + \int X_n(k)\beta(t,k)\ dk + \varepsilon_n(t)$$

The autoregressive operator Φ and its regressive equivalent B are in practice substituted by the following additive approximation using the principal component decomposition described in equation 3.2.2 (Ivanescu et al., 2015):

$$\begin{aligned} \Phi(Y_{n-1}) &= \textstyle\int Y_{n-1}(s)\varphi(t,s)\ ds \\ &\approx \textstyle\sum_{r=1}^{R} \xi_{n-1,r} \int \upsilon_r(s)\varphi(s,t)ds \\ &\approx \textstyle\sum_{r=1}^{R} \xi_{n-1,r}\tilde{\varphi}_r(t) \end{aligned} \tag{3.3.4}$$

The function $\tilde{\varphi}_r = \int \upsilon_r(s)\varphi(s,t)ds$ resembles the EFPCs used to estimate Y_{n-1} and is fitted to the curve by using the coefficient $\xi_{n-1,r}$ (cf. 3.2.2). This has to be done for all EFPCs for every input curve (both lagged curves for autoregression as well as external predictor curves).

3.4 Combined model approach

Based on the above described theory a methodology was designed to construct a model for predicting water temperatures in lagoons using the software R. ARIMAX models use discrete data as model input and aim to develop a model which will have discrete data as output. Since the data used for this work

has a high resolution, a large number of prediction steps is required to gain any relevant knowledge. Assuming a sampling interval of thirty minutes, 48 prediction steps into the future are necessary to predict as little as one day ahead. A peculiarity of ARIMAX-models is that on the long run they go towards the mean-value of their underlying data. Therefore the predictions gained from this method may be very accurate for the first few hours, but quickly afterwards loose their predictive power.
FARX models, on the other hand, are well suited to display daily temperature fluctuations, but the models may not be able to capture abrupt changes of the daily mean temperature.
To avoid the downsides of both modelling approaches mentioned above and instead make use of their strengths, a combination of them was developed. An ARIMAX model was used to forecast the daily mean temperature while a FARX model of the de-meaned time series was used to predict half-hourly fluctuations around the daily mean temperature. In the following this procedure is described in two independent steps.

3.4.1 Modelling mean water temperature

As previously described in section 2, ARIMAX models and their partial versions (most commonly AR, MA or ARMA models) are widely used in many scientific disciplines to solve a multitude of problems (e.g. for food production (Bratina and Faganel, 2008), tourism (Akal, 2004) or glaciology (Liu et al., 2015)). In the following subsection an autoregressive, integrated moving average model (ARIMAX(p,d,q)) is introduced, which is suitable to model and forecast the development of the daily mean water temperature while also incorporating the air temperature and salinity as external predictors.

An arithmetic mean of the respective data sets, computed for each day n, transforms the observed water temperature into a new time series of length n featuring one mean water temperature w_n value per day:

$$w_n = \overline{W}_n = \frac{1}{48}\sum_{t=1}^{48} W_n(t) \tag{3.4.1}$$

This is also done for the time series for the air temperature (A_n) and salinity (S_n) within the to be examined time frame. For simplicities sake the daily mean values will be from now on referred to as s_n, a_n and w_n (salinity, air and water temperature).

ARMA models assume stationarity which might be violated for specific time windows. If that is the case the data has to be once differentiated in order to account for a changing mean (therefore $d = 1$). Instead of processing the absolute values, only the difference between neighbouring entries is considered, so the change in temperature or salinity between one day and the next is now regarded. This operation can be written as: $w'_n = w_n - w_{n-1}$) with w_n being the original time series (daily mean water temperature) and w'_n being the change of the temperature. As a side effect the time series length is reduced by one. For further information on stationarity and differentiating see Hyndman and

Athanasopoulos (2018).

The following equation describes the used ARIMAX model, which is once differentiated and includes external variables:

$$w'_n = \sum_{i=1}^{p} \phi_i w'_{n-i} \ + \ \beta_1 a_n \ + \ \beta_2 s_n \ + \ \varepsilon_n \ + \ \sum_{j=1}^{q} \eta \varepsilon_{n-j} \tag{3.4.2}$$

Here we use a ARIMAX(p,q) model with two external predictors: the current air temperature (a_n) and the current salinity (s_n) (compare section 3.1). Please note that it is assumed that the data is not seasonal since the time windows are too small to display annual fluctuations. Other periodicities such as the tidal cycle are considered to have a small impact and can be neglected.

The R functions `auto.arima` and `arima` from the packages `forecast` (see Hyndman and Khandakar (2008) and Hyndman et al. (2020)) and `stats` (R Core Team, 2020), are used to compute a generic ARIMAX model. `auto.arima` determines the order (thus amount of parameters) of the best fitting ARIMAX model, selecting the one with the lowest BIC (see section 3.6: Model selection) and returning the values for p and q. Stationarity is ensured by using the Augmented Dickey-Fuller test to check for a unit root. This statistical test determines whether a data set is stationary or further differentiating is necessary (Said and Dickey, 1984).
The thus determined amount of parameters to be included into the ARIMAX model, described in equation 3.4.2, are directly passed on to the `arima` function which then estimates the parameters based on the given training data sets. The function uses maximum likelihood for determining the best fitting parameters. The overall mean is by default subtracted for the estimation and again added later on to the model output.

Now the modelled water temperature can be compared to the observed values (or strictly speaking to the mean values computed from the observations). A visual analysis of the residuals (differences between the observed values and their respective modelled equivalents) serves to verify the homoschedasticity of the errors. It is given when the variance of the residuals is constant. Heteroschedasticity on the other hand would indicate the presence of systematic errors (e.g. unaccounted seasonality) which can be visible in the residuals.
A boxplot of the residuals shows their mean, quantiles and outliers. Should the mean be significantly different from zero or the amount of outliers be high, the model should be double-checked as these can also be indicators for systematic errors. Figure 3.4 shows an exemplary plot of residuals and boxplot.

Should the computed model be free from systematic errors then the goodness of fit can be determined using the indices described in section 3.5. These measures can be used to describe their absolute fitting quality (e.g. R^2 or *RMSE*) as well as compare different model fits (e.g. *AIC* or *BIC*, cf. section 3.6).

The above mentioned procedure serves to determine a model and estimate its parameters in order to best fit the observation within the data set used for training. The developed model is then used to predict a certain amount of future

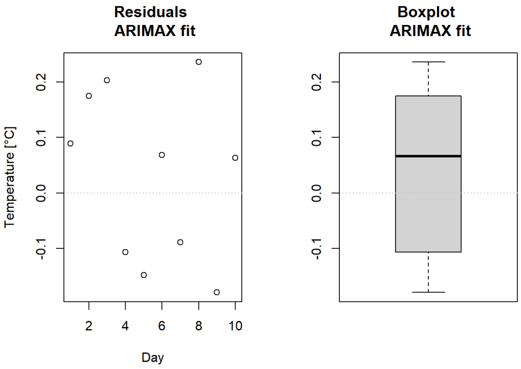


Figure 3.4: Plot of residuals and corresponding boxplot of an ARIMAX fit to water temperature data derived from the Marinetta lagoon in July 2015 (cf. chapter 4 below). Neither does the plot of residuals show any apparent pattern, nor is the distribution of the quantiles and mean far from the expected value (0), which would both indicate a systematic error. Therefore it can be assumed, that the model fit is suitable.

values which can be compared to the same amount of actual observations in the future (here ”future” refers to the time after the last observation within the training data set). For this purpose the test data set is used.

The prediction is compared to the observation, as described above for the model fit, using a residual analysis and calculating goodness of fit parameters. To quantify the model performance the sum of squared residuals (*SSR*), the coefficient of determination (R^2), the root mean squared error as well as its normalised variant (*(N)RMSE*) are computed. Should the predictions be sufficiently accurate then the model can be passed on to be used in combination with the FARX models described below.

3.4.2 Modelling daily water temperature fluctuations

The second model that will be used here is a FARX model as introduced in section 3.3. To avoid the problems described at the beginning of this section, the modelling input will be the observed data minus the mean values computed and forecast in section 3.4.1 above. This procedure is carried out for the water temperature time series and optionally can also be applied to the salinity and air temperature time series. As already done previously the fitting process for the model will be carried out solely using the training data set and then verified with the test data set.

The R package `refund` (Goldsmith et al., 2020) provides several functions for creating, handling and analysing functional data, especially for computing regression for functional data. A penalized flexible function regression can be implemented using the function `pffr`. As input the model formula for which the parameters should be estimated is needed. Functional regression terms (in this case all three regression inputs) need to be provided as function-on-function regression term with an integral operator (see section 3.3). This can be done using the function `ffpc` which regards the time series of discrete observations over n days as a set of n observed curves. The curves are then decomposed into their functional principal components using `fpca.sc`, which returns the curves as a sum of such components along with the respective estimates. Corresponding to equation 3.2.2 this can be approximated by:

$$W_n(t) \approx \sum_{r=1}^{R} \xi_{nr} \upsilon_r(t) \tag{3.4.3}$$

Here the water temperature W_n is used as an example, the same principle applies to A_n and S_n. υ_r are the basis functions, ξ the corresponding coefficients and R the amount of principal components. With the integral operator (here Φ) introduced in equation 3.3.2 this becomes:

$$\Phi(W_n(t)) = \int W_n(s)\ \varphi(t,s)\ ds = \sum_{r=1}^{R} \xi_{nr} \int \upsilon_r(s)\varphi(s,t)ds = \sum_{r=1}^{R} \xi_{nr}\tilde{\varphi}_r(t) \tag{3.4.4}$$

The full FARX model which is fitted can be written as:

$$W_n(t) = \int W_{n-1}(s)\ \varphi(t,s)\ ds + \int A_n(s)\ \beta(t,s)\ ds + \int S_{n-1}(s)\ \gamma(t,s)\ ds + \varepsilon_n(t) \tag{3.4.5}$$

This includes the term $\varepsilon_n(t)$, which is a white noise error like specified in section 3.1. In practice this equation is approximated by the function `pffr` as an additive model (see equation 3.3.4):

$$W_n(t) \approx \sum_{r=1}^{R_w} \xi_{n-1,r}\tilde{\varphi}_r(t) + \sum_{l=1}^{R_a} \xi_{n-1,l}\tilde{\beta}_l(t) + \sum_{m=1}^{R_s} \xi_{n-1,m}\tilde{\gamma}_m(t) + \tilde{\varepsilon}_n(t) \tag{3.4.6}$$

$\tilde{\varepsilon}_n(t)$ again denotes a white noise error. The estimation method is by default set to use REML (restricted maximum likelihood) which is a bias-corrected version of the maximum likelihood estimation (Dodge and Commenges, 2006), assuming that $\tilde{\varepsilon}_n(t)$ is a Gaussian random variable.

After obtaining the best fitting model for the observed training set, the residuals are computed and analysed as before (check of homoschedasticity and boxplots) and the goodness of fit of the model calculated.

Since the model uses lagged values from the previous day the `predict` function is only able to forecast one day into the future (Hyndman et al., 2020; Hyndman and Khandakar, 2008).
Also in order to use salinity as predictor, a model for forecasting the same is needed. For convenience this study simply uses a null-model which assumes that the salinity pattern of the last observed day persists and is re-used for the subsequent days. Obviously this assumptions reduces the usefulness of this predictor compared to a more advanced model.

Combining the models: The forecasts from both models for the test period were then joined, assuming and additive structure. While the ARIMAX model provides the daily mean water temperature, the inner-daily water temperature fluctuations are given by the FARX prediction. The resulting joined forecast can be compared to the actual observations within that period, a residual analysis performed and the goodness of fit determined by computing various parameters (as described above).

3.5 Indicators for goodness of fit

To be able to compare several models with each other and quantify their modelling performance it is important to use suitable indicators for the goodness of fit. Some existing methods need to be redefined in order to be applicable to functional data, while others can be used with little or no modification. In practice, the continuous functional objects, which are returned by FARX models, have to be discretised again, before they can be compared to the (discrete) observations.

Coefficient of determination Commonly denoted as R^2, this coefficient is widely used in regression analysis. It indicates how much of the variation of a variable can be explained by e.g. a regression model. When applied to compare an observed time series y with T observations to a model $\hat{y}$ the following formula is used:

$$R^2 = 1 - \frac{\sum_{i=1}^{T} \hat{\varepsilon}_i^2}{\sum_{i=1}^{T} (y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{T} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{T} (y_i - \bar{y})^2}$$

Here $\hat{\varepsilon}$ denotes the residuals and $\bar{y}$ the mean of y. A R^2 value of 1 indicates a perfect fitting model in which all residuals are zero, while a value of 0 indicates the models performance to be equal to simply taking the mean $\bar{y}$. Since the fits to the individual values can theoretically be worse than the mean it is possible to get negative values for R^2.

A downside of R^2 is that it always improves when adding more explanatory variables. To account for that an adjusted version (R^2_{adj}) has been developed, which includes a penalty term for the number of regressors (ω) a model uses. In the functional context this means the sum of all principal components, which estimate the individual curves included for calculating the FARX model (here a certain amount of EFPCs for the lagged water temperature, the air temperature and the salinity, respectively).

$$R^2_{adj} = 1 - (1 - R^2)\frac{T-1}{T-\omega-1} = 1 - \frac{\sum_{i=1}^{T} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{T} (y_i - \bar{y})^2} \cdot \frac{T-1}{T-\omega-1}$$

When applied to functional data like described in section 3.2 several minor changes have to be made:

$$
\begin{aligned}
&\bar{y} \rightarrow \bar{Y} \quad \text{with} \quad \bar{Y} = \frac{\sum_{n=1}^{N} \bar{Y}_n}{N} = \frac{\sum_{n=1}^{N} \sum_{t=1}^{T} Y_n(t)}{NT} \\
&\left.\begin{array}{l} y_i \rightarrow Y_n(t) \\ \hat{y}_i \rightarrow \hat{Y}_n(t) \end{array}\right\} \quad n = \{1, \cdots, N\}, \quad t = \{1, \cdots, T\}
\end{aligned}
\tag{3.5.1}
$$

The mean $\bar{Y}$ needs to be computed over all observations within all subsets. This data splitting into subsets has to be likewise regarded for computing the residuals and the penalty term. The functional equivalent of the equation for R^2_{adj} therefore is:

$$
R^2_{adj} = 1 - \frac{\sum_{n=1}^{N} \sum_{t=1}^{T} \; (Y_n(t) - \hat{Y}_n(t))^2}{\sum_{n=1}^{N} \sum_{t=1}^{T} \; (Y_n(t) - \bar{Y})^2} \cdot \frac{(N \cdot T) - 1}{(N \cdot T) - \omega - 1} \tag{3.5.2}
$$

A continuous version does not have to be used since the observation is available as discrete data and the model output can be discretised.

Root Mean Squared Error Abbreviated as RMSE, this indicator is a measure of accuracy of a model or prediction. Like the R^2 value it uses the sum of squared residuals ($SSR = \sum_{t=1}^{T} (y_t - \hat{y}_t)^2$) between the observed ($y_t$) and the modelled data ($\hat{y}_t$) to compute how well the model fits the observation. The closer the RMSE is to zero the better the fit. It is defined as:

$$
\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T}}
$$

To make a comparison between different data sets with possibly different value ranges possible a normalised version of the *RMSE*, the *NRMSE*, can be used. Therefore a division by the mean ($\bar{y}$) has to be undertaken:

$$
\text{NRMSE} = \text{RMSE} \cdot \bar{y}^{-1} = \sqrt{\frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{T}} \cdot \bar{y}^{-1}
$$

When applying this indicator to functional data changes similar to those for the R^2 value have to be applied. Incorporating the substitutions from equation 3.5.1 above the formula can be rewritten as:

$$
\text{NRMSE}^* = \sqrt{\frac{\sum_{n=1}^{N} \sum_{t=1}^{T} (Y_n(t) - \hat{Y}_n(t))^2}{NT}} \cdot \bar{Y}^{-1} \tag{3.5.3}
$$

Additional coefficients The previous parameters both used the squared residuals for comparison. But for certain applications it might be advisable to employ different metrics.

The L^∞-distance indicates the maximum distance of the model from the observations:

$$
L^\infty = \max_t |y_t - \hat{y}_t| \tag{3.5.4}
$$

For the functional approach the maximum distance between the observed and the modelled function needs to be computed:

$$L^{*\infty} = \max_{n,t} \left| Y_n(t) - \hat{Y}_n(t) \right| \tag{3.5.5}$$

The maximum is computed within each subset and then the maximum amongst them.

The L^1-distance indicates the total distance between model and observation:

$$L^1 = \int \mid y(t) - \hat{y}(t) \mid dt \tag{3.5.6}$$

Since the functional approach also uses discrete values for observations and in practice models discrete points, the following approximation can be applied:

$$L^{*1} = \int \left| \; Y_n(t) - \hat{Y}_n(t) \; \right| dt \approx \sum_{(n,t=1)}^{(N,T)} \left| Y_n(t) - \hat{Y}_n(t) \right| \tag{3.5.7}$$

Besides indicators that consider the whole curve it is also informative to solely regard prominent points such as the minima and maxima. As can be seen for the application in section 4 the accuracy in predicting the extreme values of each subset is important to identify notable events. Therefore it is helpful to compute the difference in maxima: $\text{diff}_{max,n} = \max\limits_{t} \; |Y_n(t)| - \max\limits_{t} \; |\hat{Y}_n(t)|$. Respectively the same applies for minima.

3.6 Model selection criteria

To determine which of several models is most suitable for describing a data set, a model selection has to be performed. The models need to be compared in regard to their performance in describing the data but simultaneously should not be too complicated in order to avoid overfitting. If too many variables are being used the model matches too strongly the signature of the training data set and cannot generalise sufficiently to predict the test set. The simplest model should be preferred if the performance of two or more models is similar (see Occam's razor (Myung and Pitt, 1997)).

An information criterion compares the quality of different models in regard to the data and thereby allows to choose the best-fitting. The two most commonly used ones are the Akaike information criterion (*AIC*) and the Bayesian information criterion (*BIC*). The *AIC* consists of two terms:

$$AIC = 2\omega + T \cdot \ln(SSR/T) \tag{3.6.1}$$

The penalty term 2ω (ω being the number of estimated parameters of the model) gets bigger if the model is more complex, penalising extensive use of parameters. The second term $T \cdot \ln(SSR/T)$ (T being the total amount of data points used and SSR being the sum of squared residuals as describe above) becomes smaller the better the model explains the data (Akaike, 1974).

The BIC on the other hand has a slightly changed formula, changing the coefficient in front of ω:

$$BIC = \ln(T)\omega + T \cdot \ln(SSR/T) \tag{3.6.2}$$

The penalty term $\ln(T)\omega$ penalises the model stronger if it was trained on more data points (Wit et al., 2012). This criterion tends to be preferable when the amount of data points used to train models differs in magnitude. For both criteria the model with the lowest score should be preferred.

These criteria can be used for discrete as well as functional data models. For the latter all (discrete) data points should be used for T. The value for ω consists of the EFPCs used for estimating the function and the modelling parameters of the FARX model.

4 Application

The approach outlined in the previous chapter was applied to the modelling of the daily pattern of water temperature in two Northern Adriatic lagoons, namely the Lagoon of Venice and the Marinetta lagoon. In this chapter, the available data set and the implementation of the methodology are described.

4.1 Lagoons in the Northern Adriatic

Both in terms of productivity and population the Veneto region in northeastern Italy can be regarded as one of the country's most important parts. With an area of approximately 550 km^2, the Lagoon of Venice is the biggest in Italy. It is located on the western shore of the Northern Adriatic sea and prominently home of the city of Venice (Canu et al., 2003). This makes it unique, since the city is both a cultural as well as an economic hotspot, inhabited by 55,000 people just on the main island and additionally hosting 25 million tourists every year. This plethora of human activity puts large stress on the local ecosystem. In addition, heavy storm surges, paired with the continuous subsidence of the sedimental underground, pose a constant threat for the survival of the city and its cultural heritage (Carbognin et al., 1995).

The lagoon's ecosystem is subject to off- and onshore influences, having three artificially stabilised inlets to the open ocean at Lido, Malamocco and Chioggia. Since major rivers have been diverted centuries ago, freshwater input comes mostly from little streams, surface run-off and precipitation. Remarkable is its shallow average depth of approximately 1 meter with some deep channels ranging up to 30 meters in depth. Tidal changes in the sea level cause frequent flooding and drying in approximately 15% of the area (Umgiesser et al., 2004).

With more then 600 km length the Po river is Italy's longest stream. In a watershed basin covering most of Northern Italy precipitation is directed towards the Adriatic where it forms the Po Delta. Besides several side arms of the Po river the delta also contains six lagoons, with various degrees of connectivity to the Adriatic Sea. Since 2015 the "Parco regionale veneto del Delta del Po" is a recognised UNESCO world heritage nature reserve.

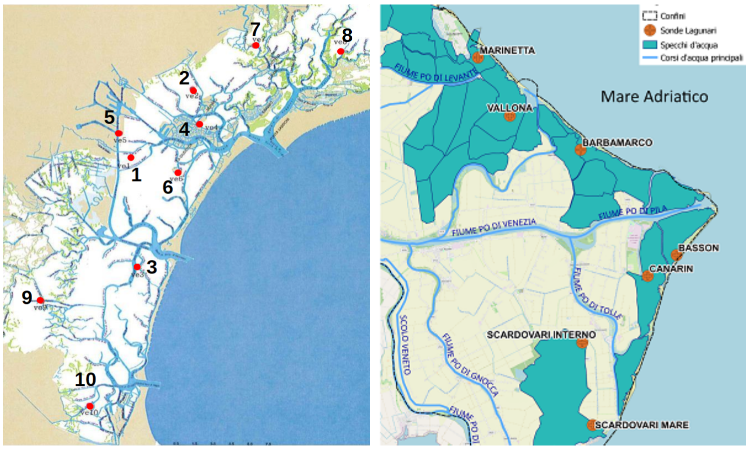
4.2 Description of measuring stations and data

The model was developed using publicly available data sets on water quality and weather conditions within the Veneto region. Since it is in the development stage and not yet designed to be applied to real-time data, model training and testing was carried out with historic data (up to 20 years old).

Water data Until 2018 the Venice Water Authority (Salvaguardia di Venezia del Magistrato alle Acque) maintained the SAMANET, a network of measuring stations within the Lagoon of Venice (Gunatilaka et al. (2009)). It consisted of ten different stations distributed in various locations within the lagoon (see figure 4.1a).

Whilst not all stations were equipped with measuring devices for every parameter, the best equipped stations measured pressure (m), water temperature ($°C$), salinity (PSU), dissolved oxygen concentration (%), chlorophyll-a concentration (g/l) and turbidity (FTU). Data was gathered every 30 minutes from January 2008 until December 2018. The number of stations and sampling frequency changed over the years: at present 5 measuring stations (1, 2, 7, 8 and 9) remain, sampling only once per hour. The Interregional administration for public works (Provveditorato Interregionale per le Opere Pubbliche per il Veneto, Trentino Alto Adige e Friuli Venezia Giulia) has assumed operation of the reduced network.
Data was provided in the framework of the research project "Venezia 2021", which is coordinated by the CoRiLa (Consorzio per il coordinamento delle Ricerche inerenti al sistema Lagunare di Venezia). The data set covers the decade 2008-2018 in which the ten sampling stations marked in figure 4.1a were collecting data every 30 minutes. The here described analysis was carried out for data between 2008 and 2018.

A similar network of measuring stations is also located within the lagoons of the Po delta, providing the same kind of data for seven different stations in six lagoons: Barbamarco, Basson, Canarin, Marinetta, Scardovari (one seawards and one landwards) and Vallona. The distribution of the measuring stations can be seen in figure 4.1b. The network is maintained by the Environmental Agency of Veneto (ARPAV - Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto).



(a) SAMANET in the Lagoon of Venice. (b) ARPAV network in the Po delta.

Figure 4.1: Maps of the networks of measuring stations for water data.

Weather data Besides the network of water measuring stations deployed in the lagoons of the Po delta, ARPAV also maintains several weather data stations in the Veneto region (see figure 4.2). From these stations two were picked for their geographical proximity to the lagoons.The station in Cavallino Treporti (160) is close to the Lagoon of Venice and at the same time on the seaside and therefore possibly suitable to detect influences determined by offshore weather conditions. The station in Porto Tolle (101) is close to the Marinetta lagoon in the Po delta, another important location for mussel farming and therefore also of interest.

The datasets feature values for the air temperature (°C) at the respective station from 1 January 2000 until 20 July 2020, continuously sampled every 15 minutes. Since the sampling frequencies were different for the two types of data sets, the weather data was adjusted for better comparability. A centred average was used, substituting the measured values with the mean of three consecutive values (centred around a half-hourly time value).

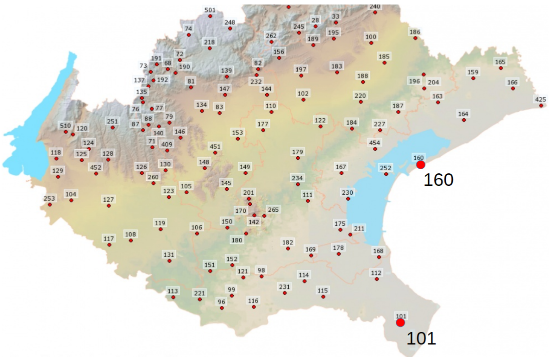


Figure 4.2: Map of the ARPA network of weather stations. The stations 101 and 160 were used for modelling.

4.3 Data quality and preprocessing

The water data features a strong data inconsistency, especially regarding data discontinuities (most likely due to measuring device failure or maintenance). Data gaps ranging from one sampling interval up to several weeks are frequent in all sets and make the data difficult to process since some of the applied R functions used for this model require the datasets to be without missing values. In order to sidestep this problem, intervals with large amounts of missing values were avoided for the model creation, or values approximated.

Since water flow rates within lagoons are usually low, it can be argued that heat convection is also low. Furthermore, based on the fact that increasing the temperature of a fixed amount of water requires a lot more energy than for air (due to the high specific heat capacity of water (see Tipler, 1999)), strong and quick peaks in the air temperature are not expected to show in the water temperature but will be spread out more evenly. Consequently the water temperature fluctuations for few missing sampling intervals can be reasonably approximated using linear interpolation. Nevertheless when applied to larger intervals this can cause strong inaccuracies, especially when modelling the inner-daily fluctuations.

The last step in the data preprocessing is splitting the data sets into two subsets each. The first set (which should at least account for approximately 80% of the data) is used for training the model, that means estimating the parameters and determining their significance. The second set (test set) is used as reference data to which the model's predictions are compared.

4.4 Functional data modelling

The methodology outlined in chapter 3 is described in detail in this section. Results are presented in the details in the following chapter (chapter 5).

4.4.1 Modelling strategy

This modelling approach aims to forecast one day of water temperature development ahead. In order to be able to do so it requires 10 consecutive days of input data for all three input variables. The external variables, however, need to be provided one day ahead (cf. table 4.1).

Table 4.1: Required model inputs to obtain a one day forecast of the water temperature (for day 11): 10 consecutive days for salinity, air and water temperature. External variables need to be provided one day ahead. The values for day 11 can originate from a model.

Variable	Sampling frequency	Days for which data is required
Air temperature	48 values/day	2-11
Salinity		2-11
Water temperature		1-10

The observed water temperature time series is modelled in two stages: 1) modelling the daily mean values using an ARIMAX model and 2) modelling the inner-daily developments by applying a FARX model. Since an additive structure is assumed, the mean values modelled in stage 1) can be deducted from the observed time series to obtain a trend-reduced time series displaying the difference from the daily mean water temperature. This is used for stage 2).

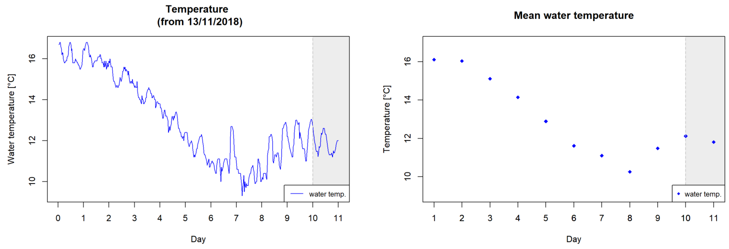
In both modelling stages several different model variants are fitted to the training data set and their goodness of fit calculated. The best-scoring variant (in terms of *BIC*, cf. 3.6) is then used for forecasting.

For both partial models a one day forecast is estimated and subsequently summed up to for the overall estimation of the detailed water temperature development for the following day. In every modelling step the intermediate results are examined for systematic errors by analysing the corresponding plot of residuals and boxplot.

4.4.2 ARIMAX model for predicting mean

The first modelling step consists of fitting an ARIMAX model to the daily mean water temperature in order to account for the daily trend. For this purpose the arithmetic mean was computed for all observations in one day respectively and the resulting time series used for modelling. The same procedure was applied for the observed air temperature and salinity.

Figure 4.3 shows an exemplary plot of the input water temperature data and the resulting mean water temperature time series.



(a) Observed water temperature time series

(b) Daily mean water temperature time series

Figure 4.3: Example of a water temperature time series used as model input. The grey background indicates where the data set is split into training and test set for cross-validation. Data was obtained from the Marinetta Lagoon in November 2018.

Six different types of ARIMAX models were applied and their fitting performances compared: ARMA, ARMAX(a), ARMAX(a,s), ARIMA, ARIMAX(a) and ARIMAX(a,s). The 'I' indicates a differentiation of the input data prior to modelling and a subsequent integration of the results (cf. chapter 3.1). Here only zero- and first-order differentiation ($d = 0, 1$) were regarded. The 'X' stand for the use of one or two external predictors, with the "(a)" indicating the use of the corresponding air temperature as external predictor while "(a,s)" indicates that both air temperature and salinity were added as external predictors.
The model orders (p and q) of each potential model are determined by the function `auto.arima` which uses the *BIC* criterion (see equation 3.6.2) to evaluate the quality of a model fit. Once the optimal order is known, it is used to fit the model to the respective training data set (10 consecutive days). The models are then evaluated based on their fitting performance and simplicity using the *RMSE* and *BIC* values for each model. The former quantifies the absolute goodness of fit in scale to the observed unit, the latter enables comparison between models with different amount of variables. The model with the lowest *BIC* score is selected and used for predicting the next day's mean temperature.
An analysis of the residuals between observation and fitted values serves to check whether there are patterns visible in the modelling errors which can indicate systematic errors.
When no such errors are evident the model can be used for forecasting. Figure 4.4 shows an exemplary ARIMAX model (fit and prediction) to the above shown data set.

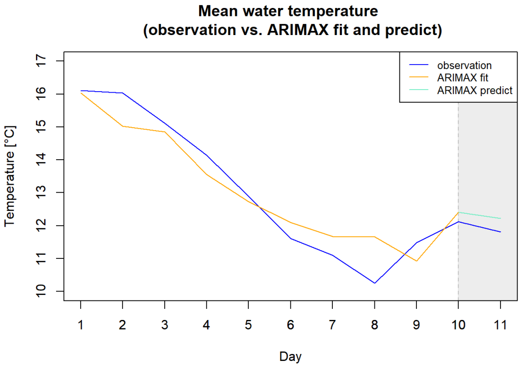


Figure 4.4: ARIMAX fit (orange) and prediction (aquamarine) to the observed mean water temperature (blue) for 10+1 days. The grey background indicates the test set for which a prediction took place.

4.4.3 FARX model for predicting daily water temperature pattern

As a second step, a FARX model is applied to predict the inner-daily fluctuations around the mean temperature. The observed values minus the fitted mean temperatures is taken as new modelling input.
Analogue to the ARIMAX model before, several different types of FARX models were tested and evaluated in order to determine which performs the best. Three model structure were tested, namely FAR, FARX(a) (with air temperature as predictor) and FARX(a,s) (with both air temperature and salinity as predictors). The data for the two external input variables can be provided in two variants: 1) unprocessed ("real") or 2) trend-reduced ("detrended") by deducting the respective daily mean values. This option is also tested.
The designated best model is then applied to the detrended training set by using the function `pffr` which fits the one day lagged water temperature and the current air temperature (and optionally also the salinity) as functional elements to the data. In order to use the time series as functional elements they have to be approximated as a sum of principal components using `ffpc` and `fpca.sc` (see Chapter 3.4.2). The threshold of variability which should be explained by the estimated functional principal components is set to 99% (see also Chapter 3.2). Each principal component of each input variable is estimated using the restricted maximum likelihood and the resulting fitted model returned.

After the model parameters are estimated, the respective models are applied for forecasting the inner-daily fluctuations for the next day. Again the air temperature is assumed to be known while for the salinity a null-model is applied which assumes that the next day is identical to the previous day. This approximation has to be made since its not realistic to assume a salinity model with such a high resolution will be available for application under real conditions. Figure 4.5 shows an exemplary FARX model fit and prediction to the above shown input data.

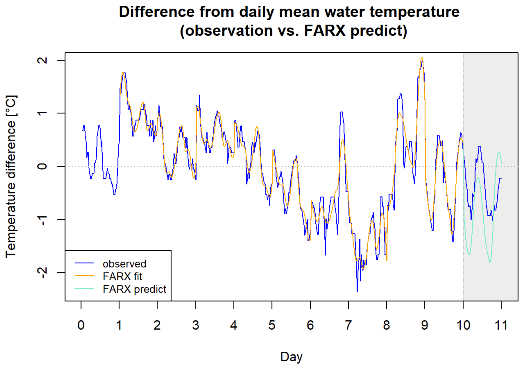


Figure 4.5: FARX fit (orange) and prediction (aquamarine) to the demeaned time series of the observed water temperature (blue). The grey background indicates the test set for which a prediction took place.

4.4.4 Combined temperature model

So far two models have been implemented: an ARIMAX model suitable for modelling and forecasting the daily mean water temperature and a FARX model suitable for forecasting the inner-daily temperature fluctuations. Combined, these two predictions add up to one model which is able to predict the water temperature with very high resolution (30 min), as was the desired outcome. Figure 4.6 shows an example of the application of the full model to the above shown water temperature time series.

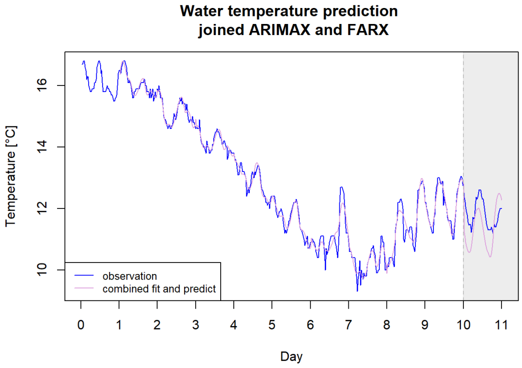


Figure 4.6: The observed water temperature (blue) is modelled and predicted by the combined model (magenta). The grey background indicates the test set for which a prediction took place.

4.5 Selection of time periods and measurement stations

Naturally summer and winter are of special interest for aquacultural modelling since their extreme temperatures and weather events can often cause conditions that directly impair the stock health or anticipated yield (De Silva and Soto,

2009).
Intense and prolonged heat waves in summer can cause warm water temperatures and high salinity in lagoons, whereas periods of extreme cold or precipitation can shift the water conditions towards the other end of the spectrum with cold water temperatures and low salinity (the latter is connected to reduced evaporation coupled with increased local precipitation and river discharge coming from mountainous regions).

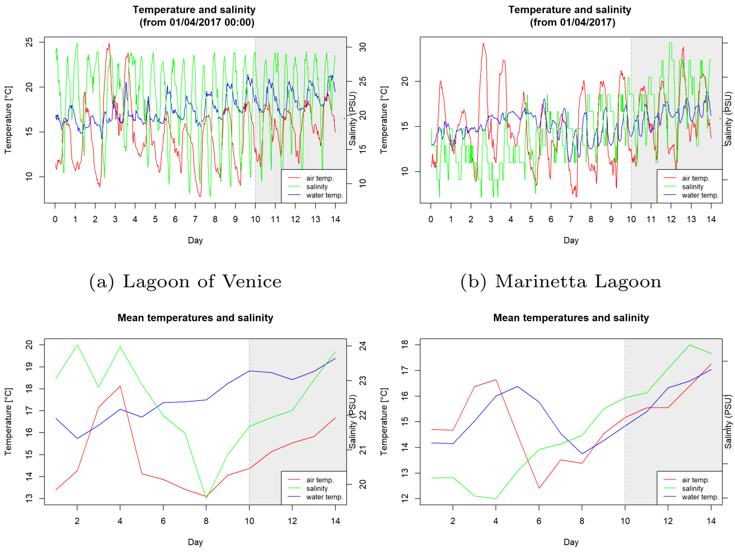
The heat wave which occurred in June and July 2015 in Europe (Russo et al., 2015) serves as an example period in which temperatures were critically high. It lasted from the end of June almost until the end of July. The period from 10 until 19 July is chosen for training the models. It exhibits constantly high temperatures often exceeding $30°C$ (both air and water temperature). At the same time the data quality is good with very few missing values that can be interpolated without causing significant modelling errors. These modelling periods will furthermore be referred to as 'V_summer' and 'M_summer'.
The storm Adrian, which made landfall in Italy in the end of October 2018, brought a series of heavy and sustained precipitation events, storm surges and strong winds, causing severe damage in Northern Italy and neighbouring states (see for example BBC News (2018)). The period around these events serves as an example for critical autumn/winter conditions in the examined water bodies. For reasons not further known the data sets from the Marinetta lagoon are of exceptionally poor quality during that time with large periods of missing data. Thus limited in choice the time from 13 until 22 November is used as training set. These periods are from now on referred to as 'V_autumn' and 'M_autumn'. To also account for spring weather conditions the period from 1 to 10 April 2017 was randomly selected as the third training period. Following the previous pattern these time frames are called 'V_spring' and 'M_spring'.

For the Lagoon of Venice station 7 of the SAMANET network (see figure 4.1a) is used to provide data on water temperature and salinity. It is the northernmost measuring station located just northwest of the Burano island. Its relative isolation from shipping routes, deep channels and urban agglomerations (such as the city of Venice and the port of Marghera) make it a relatively secluded area with little marinal influence which is potentially suitable for aquaculture. Data of the air temperature is taken from the observations made by station 160 of the ARPA network (see figure 4.2) as it is closest to the lagoon and the water measuring station (with the exception of station 252 in the heart of the city of Venice).

Within the Marinetta Lagoon only one measuring station for water data is deployed (see figure 4.1b). For air temperature data the station 101 of the ARPA network (see figure 4.2) is used. Station 112 seems equally suitable (proximity to coast and water measuring station).

Figures 4.7, 4.8 and 4.9 show the observed data input and the resulting mean time series for the three periods in the lagoons of Venice and Marinetta respectively.
For the analysis of the individual cases the same R script is used, showing the high flexibility of this method. In order to change the location and time period,

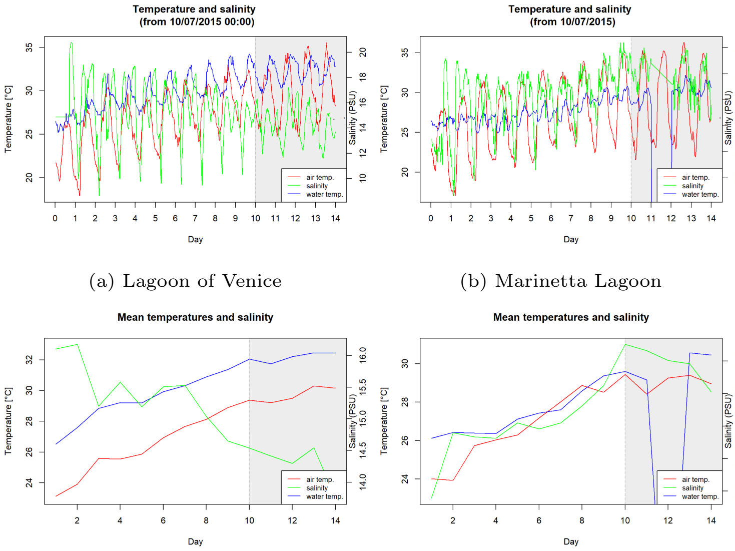
only very few settings have to be adjusted.



(a) Lagoon of Venice

(b) Marinetta Lagoon

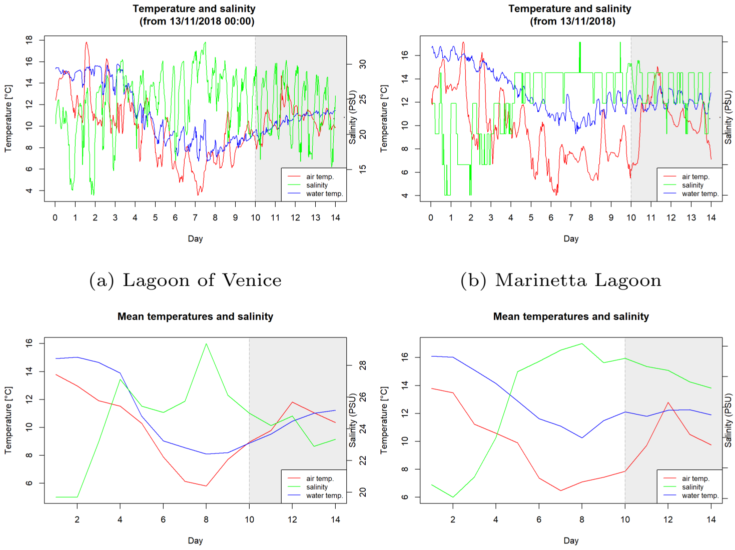
Figure 4.7: Observation (top) and daily mean values (bottom) for the both lagoons in spring. Red indicates the air temperature, blue the water temperature and green the salinity.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 4.8: Observation (top) and daily mean values (bottom) for both lagoons in summer. Red indicates the air temperature, blue the water temperature and green the salinity.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 4.9: Observation (top) and daily mean values (bottom) for both lagoons in autumn. Red indicates the air temperature, blue the water temperature and green the salinity.

5 Results

In chapter 4.5 two lagoons and three time periods of interest were identified. To each of these six data sets the above described modelling procedure was applied. The findings are presented in this chapter.

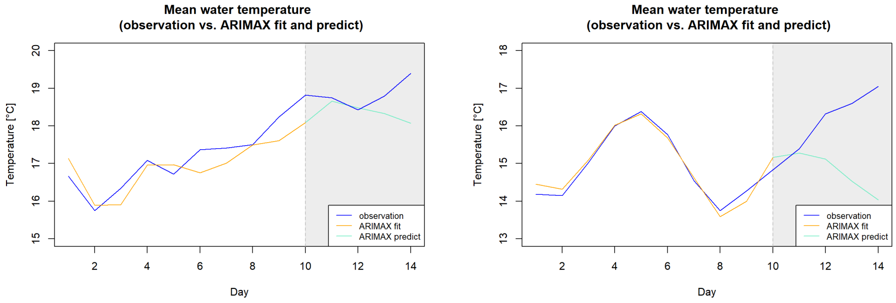
5.1 ARIMAX model

To ensure that the optimal ARIMAX model is applied, six variants were applied to fit the training set of each data set. The fitting performance was evaluated based on the respective *BIC* score. Table 5.1 shows the *BIC* scores for all potential ARIMAX models. The cells which are highlighted green are the models which are being used for forecasting. Please note the two exceptions for M_spring and M_autumn in which the model with the lowest *BIC* is not used but a different model instead. In these two cases the predictive performance of the models is so weak, that further modelling based on these results would not be viable. For the here intended proof of concept this inconvenience is condoned but for a practical implementation the use of a ARIMAX model with better predictive power would be advisable.

Table 5.1: *BIC* scores of the different models for all examined time periods (rounded to two decimal digits). The green cells indicate the models which were used for forecasting. The red cells indicate models which score better for fitting but perform too weak in the prediction to be applied.

	V_spring	V_summer	V_autumn	M_spring	M_summer	M_autumn
ARMA	-4.29	-9.74	4.29	-19.68	-8.79	-1.94
ARMAX(a)	-8.88	-15.55	-0.03	-22.01	-11.83	-1.86
ARMAX(a,s)	-11.60	-14.49	-5.34	-27.26	-29.07	-3.58
ARIMA	-2.43	-3.26	4.95	-15.77	-9.56	0.16
ARIMAX(a)	-9.45	-8.03	0.73	-15.79	-11.40	-1.98
ARIMAX(a,s)	-9.76	-8.14	-1.24	-31.73	-16.71	-12.09

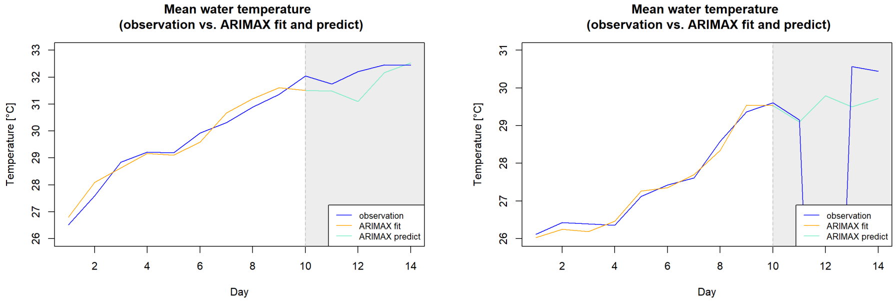
After the optimal ARIMAX model was determined it can be applied to forecast the mean water temperature for the following days. Figures 5.1, 5.2 and 5.3 below show the model fit to the training data and the forecasts for the next four days. For the joined model only the first day forecast is of concern, therefore the sudden drop in the water temperature observed in summer in the Marinetta lagoon (cf. 5.2b) is irrelevant. It results from a measurement error.



(a) Lagoon of Venice

(b) Marinetta Lagoon

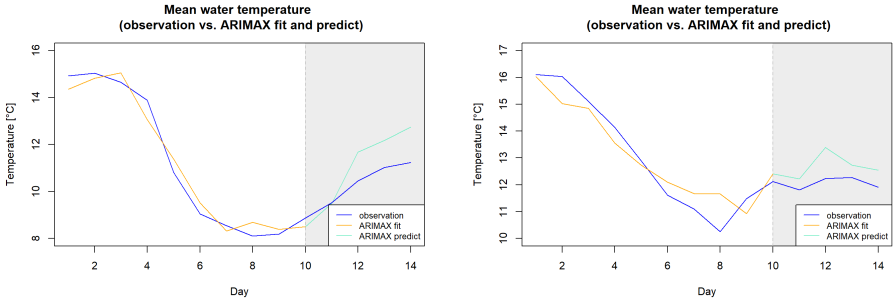
Figure 5.1: Forecast by the ARIMAX models for the Venice lagoon (left) and Marinetta lagoon (right) in spring. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next four days.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.2: Forecast by the ARIMAX models for the Venice lagoon (left) and Marinetta lagoon (right) in summer. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next four days.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.3: Forecast by the ARIMAX models for the Venice lagoon (left) and Marinetta lagoon (right) in autumn. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next four days.

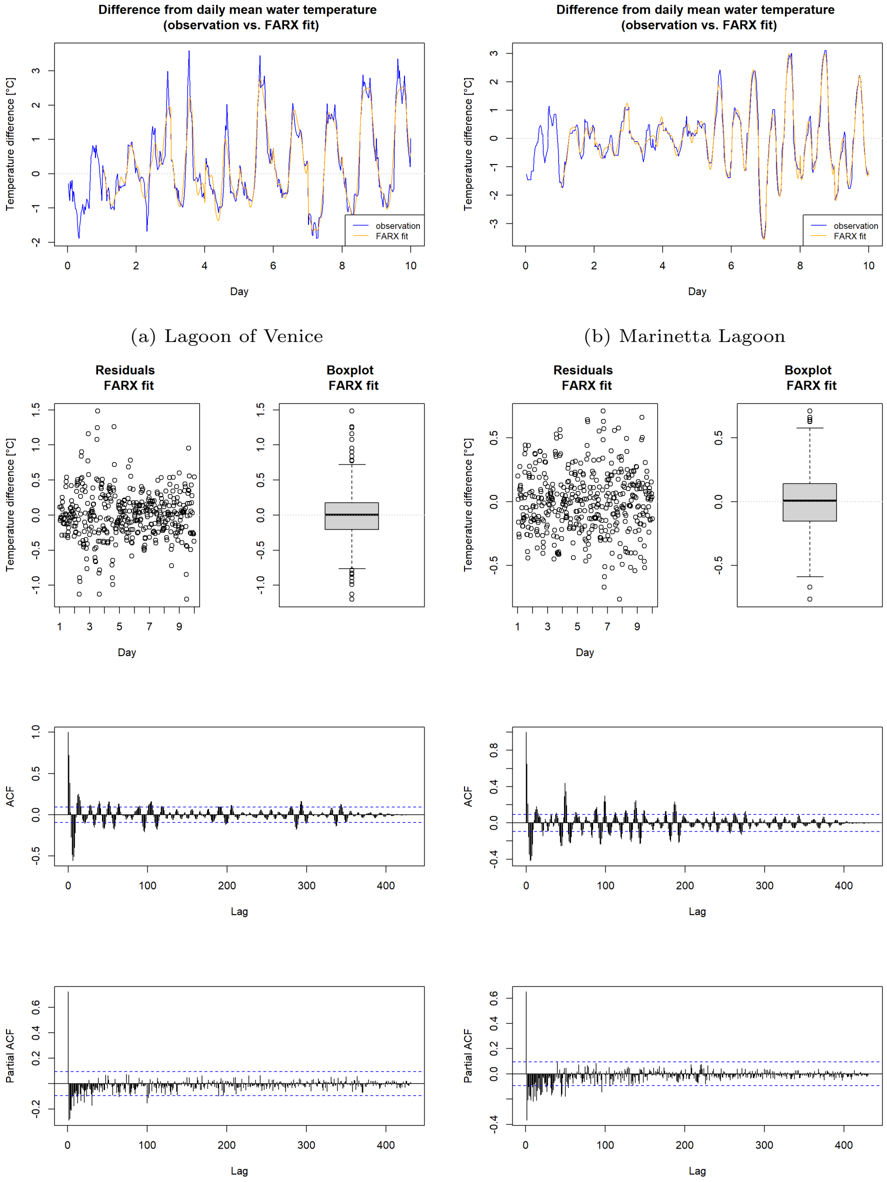
5.2 FARX model

In the second modelling step several different FARX models were fitted to the respective demeaned water temperature time series. To every set a FAR, a FARX(a) and a FARX(a,s) model was applied. Additionally the option to detrend the input time series of the two external variables (cf. chapter 4.4.3) was examined. Thus a total of twelve variants was applied to every individual case. Their fitting performance was evaluated based on the corresponding *BIC* scores. Table 5.2 shows the *BIC* scores for all the possible model combinations. Again the best scoring models are highlighted in green.

Table 5.2: *BIC* scores of the different FARX models for all examined time periods (rounded to two decimal digits). The green cells indicate the models which were used for forecasting.

		V_spring	V_summer	V_autumn	M_spring	M_summer	M_autumn
real air real salinity	FAR	-612.43	-820.19	-457.70	-864.31	-1021.18	-643.68
	FARX(a)	-852.66	-934.31	-863.38	-1162.81	-1120.47	-1274.40
	FARX(a,s)	-818.42	-900.27	-825.04	-1125.85	-1087.94	-1289.37
real air detrended salinity	FAR	-612.43	-820.19	-457.70	-864.31	-1021.18	-643.68
	FARX(a)	-852.66	-934.31	-863.38	-1162.81	-1120.47	-1274.40
	FARX(a,s)	-814.80	-899.60	-829.47	-1132.01	-1095.67	-1294.40
detrended air real salinity	FAR	-612.43	-820.19	-457.70	-864.31	-1021.18	-643.68
	FARX(a)	-845.29	-934.08	-864.71	-1163.23	-1120.35	-1275.24
	FARX(a,s)	-820.98	-903.13	-838.97	-1127.29	-1099.17	-1301.18
detrended air detrended salinity	FAR	-612.43	-820.19	-457.70	-864.31	-1021.18	-643.68
	FARX(a)	-658.94	-833.84	-628.36	-1167.12	-1109.98	-1217.91
	FARX(a,s)	-827.29	-926.88	-815.79	-1143.89	-1101.03	-1301.92

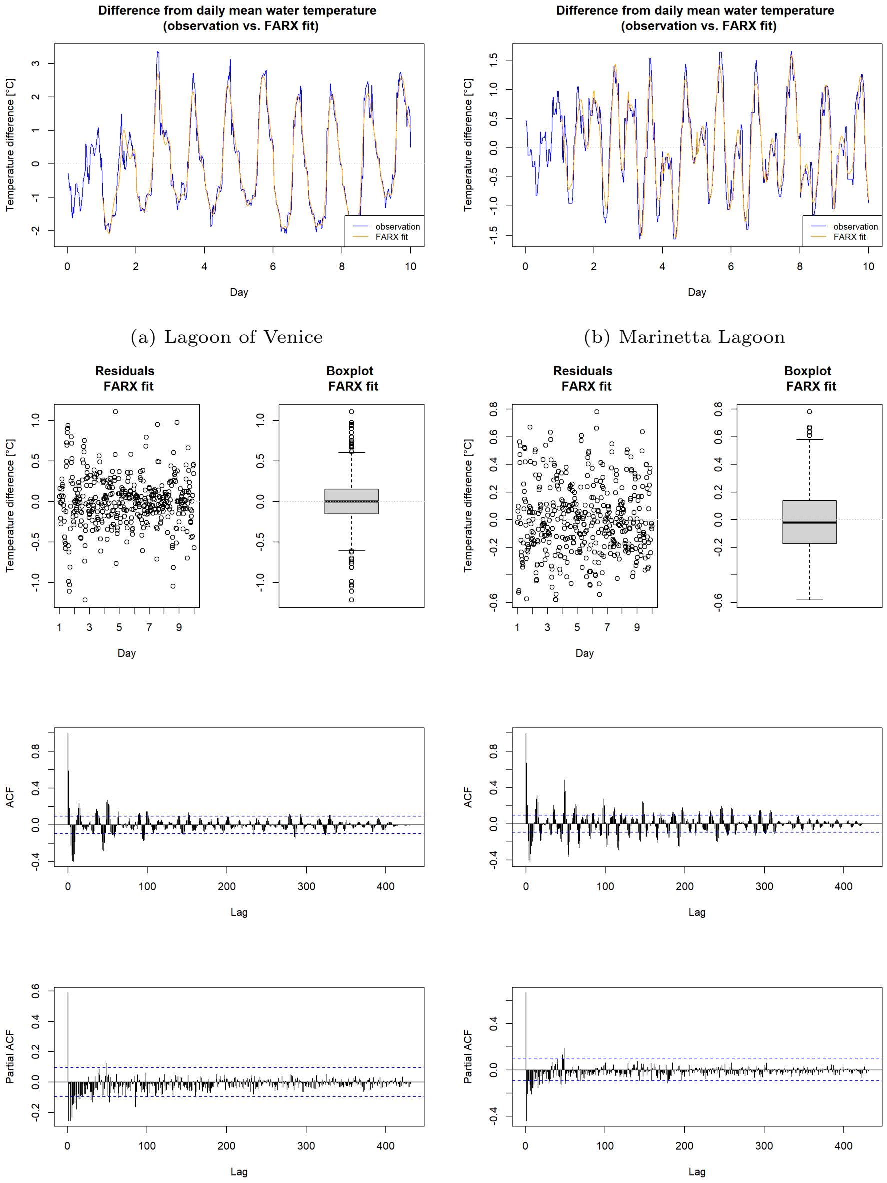
The best scoring FARX model variant was chosen and a visual analysis of the plot of residuals and the boxplot carried out to ensure that there are no systematic errors. Figures 5.4, 5.5 and 5.6 show the respective fitted FARX model to the detrended data, the corresponding residual analysis as well as the (partial) correlograms of the residuals. In all cases the residual analysis suggests that the residuals of the model fits are distributed evenly around zero.



(a) Lagoon of Venice

(b) Marinetta Lagoon

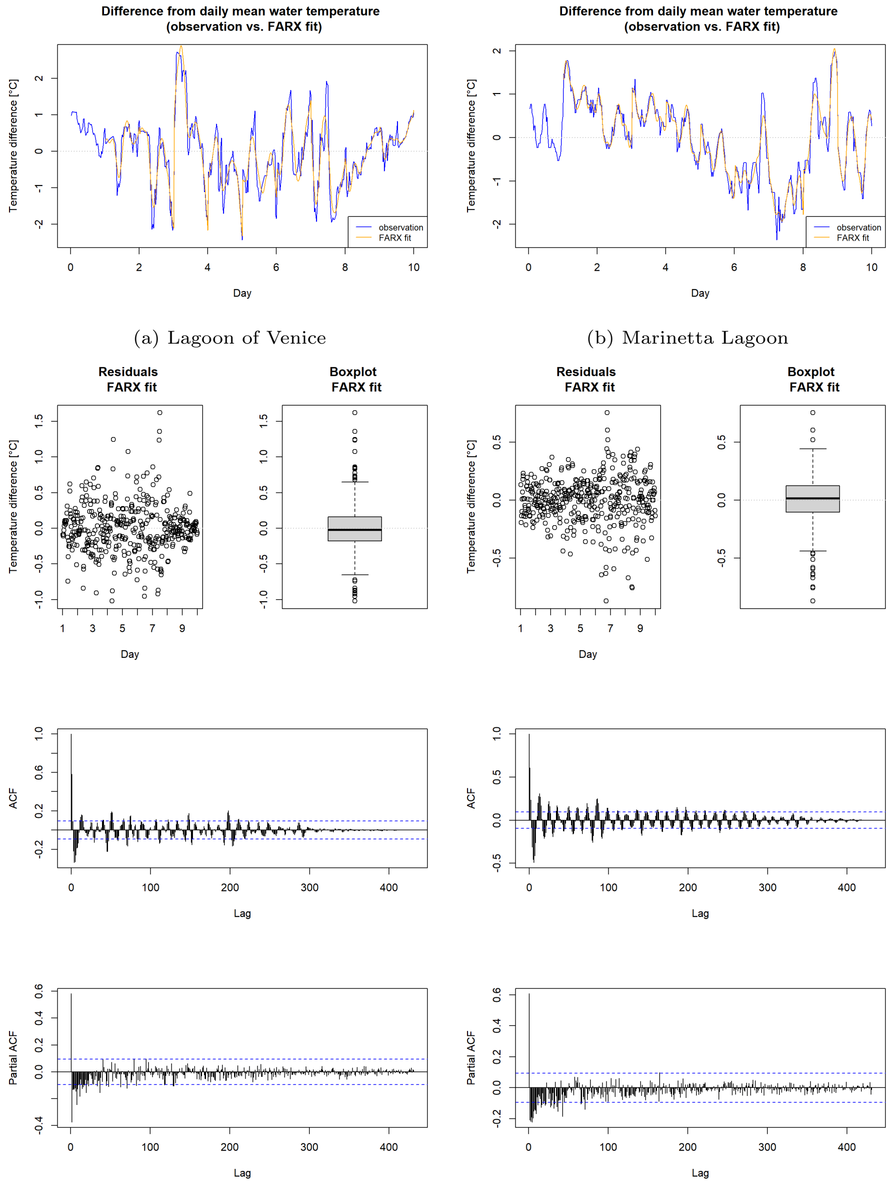
Figure 5.4: Top: Observation (blue) and FARX fit (orange) of the demeaned data for both lagoons in spring. Blue indicates the observed data while orange depicts the estimated model fit. Middle: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors. Bottom: Autocorrelation function and partial autocorrelation function of the residuals.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.5: Top: Observation (blue) and FARX fit (orange) of the demeaned data for both lagoons in summer. Blue indicates the observed data while orange depicts the estimated model fit. Middle: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors. Bottom: Autocorrelation function and partial autocorrelation function of the residuals.

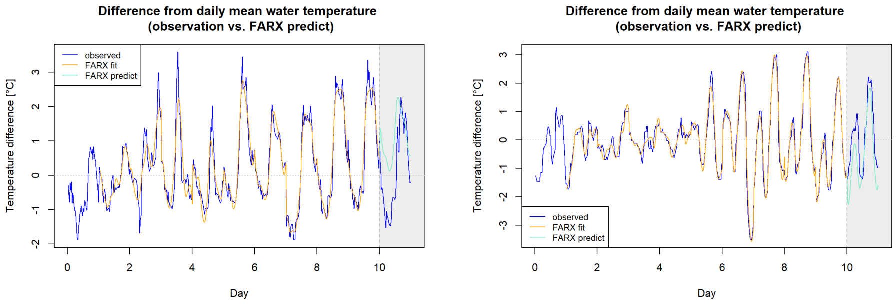


(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.6: Top: Observation (blue) and FARX fit (orange) of the demeaned data for both lagoons autumn. Blue indicates the observed data while orange depicts the estimated model fit. Middle: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors. Bottom: Autocorrelation function and partial autocorrelation function of the residuals.

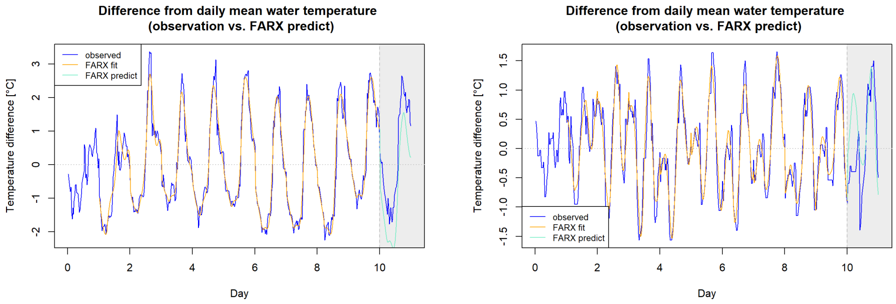
Afterwards the FARX models were applied for forecasting one day into the future. Figure 5.7, 5.8 and 5.9 shows the respective FARX forecasts for the Lagoon of Venice (left) and the Marinetta Lagoon (right).



(a) Lagoon of Venice

(b) Marinetta Lagoon

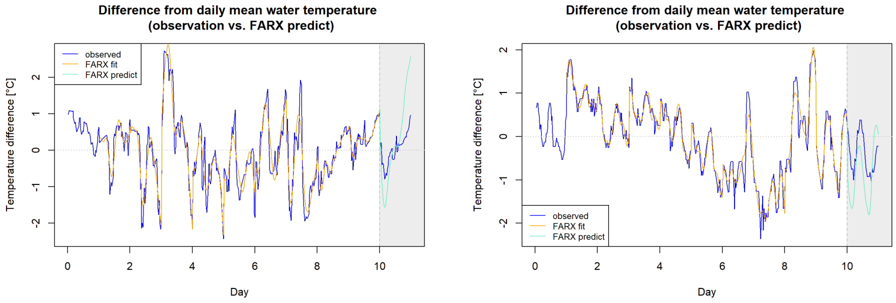
Figure 5.7: Forecast by the FARX models for both lagoons in spring. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next day.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.8: Forecast by the FARX models for both lagoons in summer. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next day.

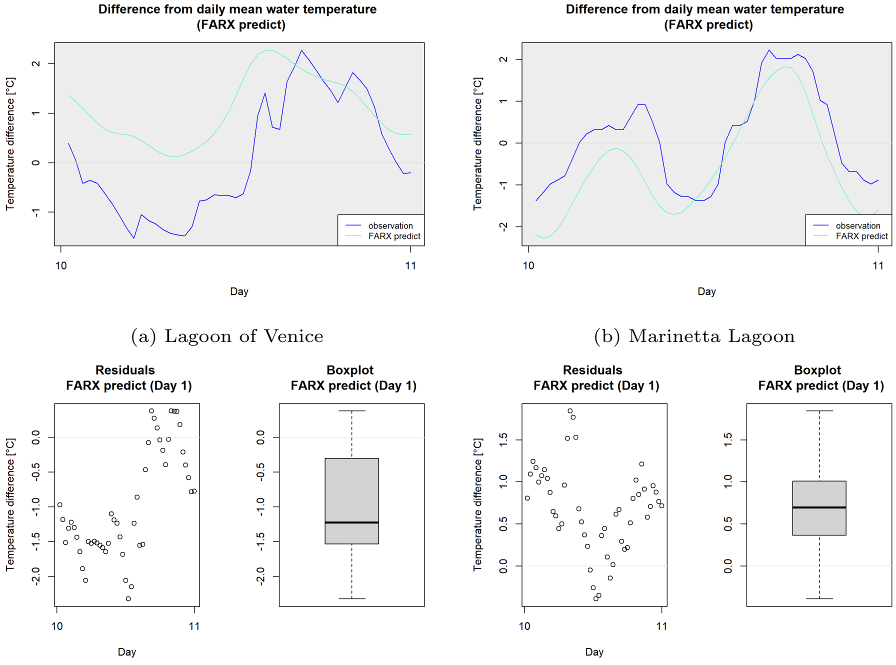


(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.9: Forecast by the FARX models for both lagoons in autumn. Blue indicates the water temperature, orange the fit to the training set and light blue the forecast for the next day.

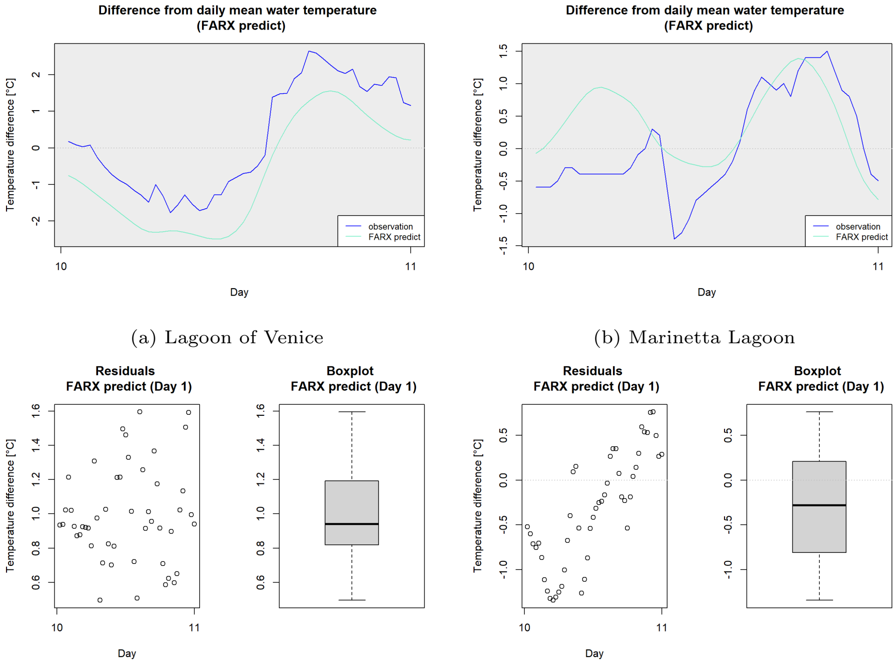
Figures 5.10, 5.11 and 5.12 show a detailed view of the predicted day (top) and the corresponding residual analysis with plot of residuals and boxplot (bottom). Especially noteworthy is the residual pattern of for the Lagoon of Venice in autumn as shown in figure 5.12a, which displays a non-random linear pattern. At the beginning of the day the prediction is too low, at the end of the day too high (by more than $1.5°C$).



(a) Lagoon of Venice

(b) Marinetta Lagoon

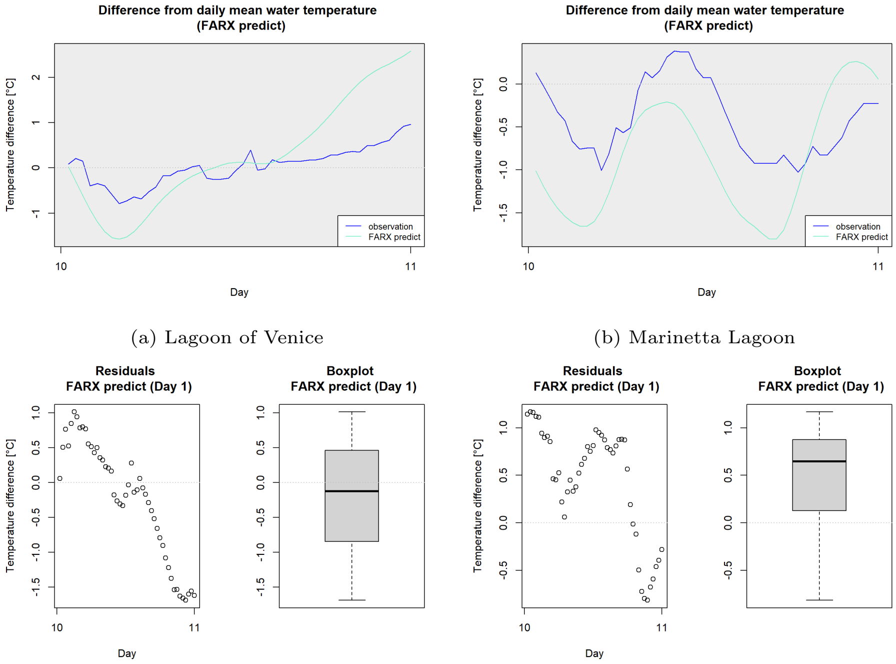
Figure 5.10: Top: Observation (blue) and FARX prediction (light blue) of the demeaned data in spring. Bottom: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors.



(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.11: Top: Observation (blue) and FARX prediction (light blue) of the demeaned data in summer. Bottom: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors.



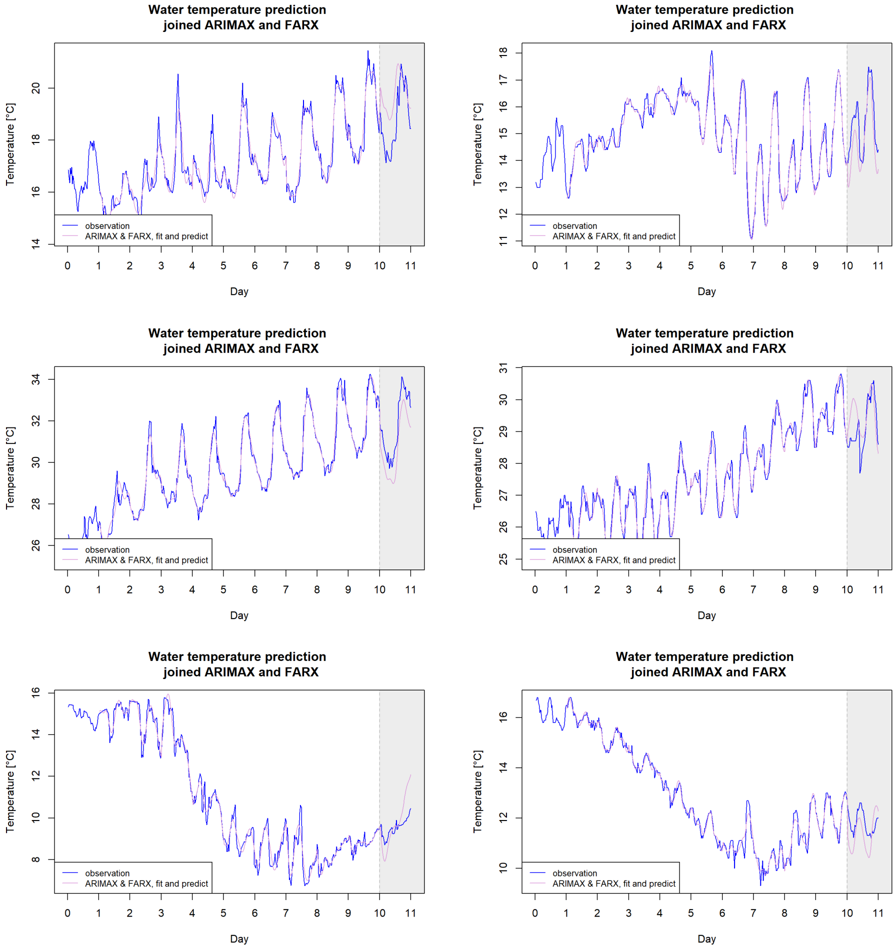
(a) Lagoon of Venice

(b) Marinetta Lagoon

Figure 5.12: Top: Observation (blue) and FARX prediction (light blue) of the demeaned data in autumn. Bottom: corresponding residual analysis with plot of residuals and boxplot to check for systematic errors.

5.3 Combined model

After both modelling steps have been carried out successfully, the individual models can be added up to form the combined model estimation. Figure 5.13 shows the resulting model fit including forecast (magenta) in comparison to the observed water temperature (blue).



(a) Lagoon of Venice

(b) Marinetta lagoon

Figure 5.13: Forecast by the combined models for the Venice lagoon (left) and Marinetta lagoon (right) in spring (top), summer (middle) and autumn (bottom). Blue indicates the observed water temperature and magenta the model output as expressed by the combination of the respective ARIMAX and FARX models.

Table 5.3 shows the respective scores for various goodness of fit indicators of the combined model in respect to the observation. The calculations only take the predicted day into account (48 observations). The following indicators are computed: sum of squared residuals (*SSR*), coefficient of determination and its adjusted version which considers the number of variables (R^2 and R^2_{adj}), the root mean squared error and its normalised variant which enables comparison between different data sets (*RMSE* and *NRMSE*), the L^1 distance as total distance between observation and prediction and lastly the L^∞ distance which indicates the maximum distance between these two curves. Depending on the envisaged application these parameters are more or less important and could even be used for model selection instead of the *BIC*.

Table 5.3: Indicators for the respective goodness of fit between observation and the one-day forecast by the combined model (rounded to two decimal digits).

	V_spring	V_summer	V_autumn	M_spring	M_summer	M_autumn
SSR	74.22	50.94	34.93	34.60	23.57	25.98
R^2	0.76	0.84	0.65	0.86	1.00	0.64
R^2_{adj}	1.19	1.24	1.16	1.25	1.16	1.35
RMSE	1.24	1.03	0.85	0.85	0.70	0.74
NRMSE	0.07	0.03	0.08	0.05	0.03	0.06
L^1	51.21	47.65	32.47	35.08	27.91	32.31
L^∞	2.32	1.60	1.69	1.84	1.34	1.17

6 Discussion

This chapter analyses and comments the results presented in chapter 5, focussing on their interpretation and methodological aspects.

6.1 ARIMAX

As far as the mean trend forecast is concerned, it can generally be observed that ARMAX(a,s) models (including both air temperature and salinity as external predictors) perform the best overall for all six examined situations. Table 5.1 shows, that this is true for the time periods V_spring, V_autumn and M_summer, in which the inclusion of air temperature and salinity increases the model performance noticeably, while the differentiation of the input data prior to modelling on the other hand does not do so.
Judging by the *BIC* scores the summer set in the Lagoon of Venice (V_summer) can best be described by an ARMAX(a) model without salinity as external input ($BIC_{\text{ARMAX(a)}} = -15.55$), but the score for the respective ARMAX(a,s) model ($BIC_{\text{ARMAX(a,s)}} = -14.49$), which takes salinity into account, is just slightly inferior (notice that the *BIC* already regards model complexity by incorporating a penalty term based on the amount of variables).
For the spring and autumn sets in the Marinetta Lagoon (M_spring and M_autumn) the ARIMAX(a,s) models (two external predictors and one differentiation step) exhibit the best *BIC* score, but when applying them for the next day's prediction, these models demonstrate weak predictive power, which exposes them as unsuitable for further use in the next modelling step. Therefore, the next best models (in both cases ARMAX(a,s)) are adopted instead. Figure 6.1 shows the mean trend forecast of both models for M_autumn. The predictive power of the ARMAX(a,s) model (left) is significantly higher than that of the ARIMAX(a,s) model (right), despite its better score. This issue is most likely attributable to an overfit of the ARIMAX(a,s) model to the training set and indicates need for further improvement of the trend forecasting methodology.

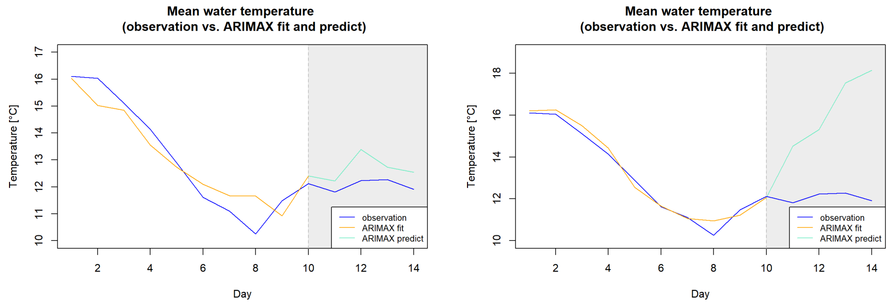


Figure 6.1: Mean trend forecast for M_autumn using an ARMAX(a,s) (left) and an ARIMAX(a,s) (right) model. Even though the ARIMAX(a,s) model has a better *BIC* score, its forecasting power is futile.

The resulting forecasts, as depicted in figure **??**, are all able to reflect the prevalent dynamic of the mean trend and return adequate values for the next-day forecast (see forecasting errors in table 6.1). This is remarkable, given the fact that these models are trained on a very small data basis (10 consecutive days), a choice which was taken deliberately, in order to test this methodology's capacity to give a precise prediction even when based on little data input. However, keeping the aspired high prediction accuracy in mind, these errors can still be severe and cause a underperformance of not only the ARIMAX model but also the joined models. These results could be improved by extending the time window for the estimation of the ARIMAX model, and/or testing other methodologies for predicting the local trend.

Table 6.1: Forecasting errors for the next-day forecast of the respective ARIMAX models (rounded to two decimal digits).

	V_spring	V_summer	V_autumn	M_spring	M_summer	M_autumn
forecasting error 1 day ARIMAX	0.09	0.26	0.04	0.12	0.06	-0.42

6.2 FARX

Unlike the mean trend models, which mostly (all but V_summer) include both air temperature and salinity as external predictors, the optimal FARX models do not include salinity. Table 5.2 shows that 5 out of the 6 best-scoring models use only air temperature as an external variable. An exception thereof is M_autumn for which the FARX(a,s) model with both predictors ranks best. This observation indicates that, while being a helpful predictor for the daily water temperature, salinity is, in these cases, not relevant for forecasting the high-frequent inner-daily temperature fluctuations. The importance of salinity is presumed to vary, depending on the selected time window (i.e. time of the year).

Furthermore, the choice of either using the originally observed time series (real) of the external inputs or their trend-reduced (detrended) variant does not seem to play a significant role for the results, as in most cases, the difference in *BIC* scores between detrended and real input is negligible (also see table 5.2). Overall, it can be said, that the real observations score slightly better, which could be due to the inaccuracies originating from the demeaning procedure.

The resulting model fits to the training set are very accurate, display the variable development in detail and neglect only some extreme spikes. The predictive power, however, is more ambiguous. As can be seen in the figures **??** and **??**, the general dynamics of the observations is resembled well by the next-day predictions, but in most cases an offset in the mean value can be observed (cf. V_spring, V_summer, M_spring, M_summer and M_autumn). While V_spring and M_summer display a decreased amplitude in regard to the observation, V_autumn demonstrates a highly increased amplitude, peaking at $1.5°C$ higher than the observation. When comparing the different input periods (e.g. relatively regular V_summer with inconsistent V_autumn), it is apparent that input data with increased irregularity is negatively correlated with the predictive power of the resulting FARX model. To overcome this, the extension of the training set could be appropriate.

6.3 Model results

The resulting joined model reflects the particular strengths and shortcomings of its FARX model, as their dynamics are identical, the only difference being the added mean value. It can be noted that the modelling performance is generally worse in spring than in summer and autumn. Likewise, the performance in the Marinetta Lagoon is better than in the Lagoon of Venice (see table 5.3). A possible reason for this could be the high complexity of the Lagoon of Venice regarding geomorphology, ecosystem linkage and anthropogenic influences.

A problem which still needs fixing in order to increase convenience and utility of such models is the yet limited forecasting horizon. As of now, the described methodology is suitable to reliably predict only one day. Prediction performance for longer periods deteriorates quickly. This is, in large part, due to the deployment of an autoregressive process of first order (FAR(1)). When predicting a second day into the future, only the prediction of the first day is used as input (and eventually external predictors), so that it can be observed that the pattern of the first day is somewhat amplified. Most often this fails to describe the observation adequately.

Overall, this modelling approach shows promising results in all six cases, demonstrating the great potential of functional data analysis for obtaining reliable forecast using a relatively small data set for selecting the model structure and estimating the parameters.

7 Conclusion

Functional data analysis is a sound data driven modelling approach, which seems promising for predicting the daily pattern of water quality variables in a highly dynamic transition water body. The results presented in chapter 5 show that reliable, 1-day ahead water temperature predictions can be obtained on the basis of the previous 10 day time series of water temperature, air temperature and salinity and on the one-day air temperature and salinity forecast. The flexible scripts coded in R for implementing the FARX model could be easily adapted to model other relevant water quality variables , such as dissolved oxygen, and salinity, which could be predicted on the basis of precipitation pattern and river discharges. Furthermore, the models could be easily implemented for updating the predictions on a regular basis, i.e. on a weekly basis, provided that the input data can be downloaded in real time.

On the other hand, in order to be used in managing halieutic resources, the prediction time horizon should be extended as far as a possible: a weekly prediction would, indeed, be desirable and 2-4 days ahead forecast would already be very useful. Of course, the reliability would decrease towards the end of this horizon, as it happens for weather forecast.
As highlighted in chapter 6, this goal can be achieved by improving both components of the model, i.e. the trend and the seasonal one, i.e. the daily pattern. As far as the trend is concerned, daily mean values could be predicted on the basis of a larger data set, using the same methodology, i.e. ARIMAX, or other local methodologies for the extrapolation of the trend could be used, e.g. polynomial fitting. The seasonal component was simulated using functional autoregressive process of first order (FAR(1)): in order to extend the prediction horizon, more functional autoregressive terms could be added, thus adopting a FAR(p) process. A different approach to solve this problem could be enlarging the length of one functional object, which is now one day, to two or more days. Therefore, one prediction step ahead would include multiple days: in this case, however, the amount of input data would increase, as the length of the training set would have to be extended substantially.

Therefore, this work can be seen as the starting point for the development of a functional modelling approach to tackle issues, in which the prediction of environmental variables is relevant. The methodology shows promising potential for predicting highly dynamic patterns and, therefore, could be used for assessing risks for organisms and ecosystems associated with the occurrence of high/low values of uncontrollable variables, driven by physical forcings. The frequency of such events, i.e. heat waves, hypoxias, is expected to increase due to climate change.
In our changing world detailed modelling and precise forecasting is becoming increasingly important to understand the mechanisms which are going to shape the next decades and how this is going to affect human life, directly and indirectly.

References

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

Akal, M. (2004). Forecasting Turkey's tourism revenues by ARMAX model. *Tourism Management*, 25(5):565–580.

BBC News (2018). Venice under water as deadly storms hit Italy. *Retreived from: https://www.bbc.com/news/world-europe-46029302.*

Benyahya, L., Caissie, D., St-Hilaire, A., Ouarda, T. B., and Bobée, B. (2007). A review of statistical water temperature models. *Canadian Water Resources Journal*, 32(3):179–192.

Bergamasco, A., Carniel, S., Pastres, R., and Pecenik, G. (1998). A unified approach to the modelling of the Venice Lagoon–Adriatic Sea ecosystem. *Estuarine, Coastal and Shelf Science*, 46(4):483–492.

Biess, A., Nagurka, M., and Flash, T. (2006). Simulating discrete and rhythmic multi-joint human arm movements by optimization of nonlinear performance indices. *Biological cybernetics*, 95(1):31–53.

Brander, K. M. (2007). Global fish production and climate change. *Proceedings of the National Academy of Sciences*, 104(50):19709–19714.

Bratina, D. and Faganel, A. (2008). Forecasting the primary demand for a beer brand using time series analysis. *Organizacija*, 41(3).

Caissie, D., El-Jabi, N., and St-Hilaire, A. (1998). Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering*, 25(2):250–260.

Canu, D. M., Solidoro, C., and Umgiesser, G. (2003). Modelling the responses of the Lagoon of Venice ecosystem to variations in physical forcings. *Ecological modelling*, 170(2-3):265–289.

Cao, J. (2019). Lecture "Functional Data Analysis". *Simon Fraser University, Vancouver.*

Capuzzo, E., Lynam, C. P., Barry, J., Stephens, D., Forster, R. M., Greenwood, N., McQuatters-Gollop, A., Silva, T., van Leeuwen, S. M., and Engelhard, G. H. (2018). A decline in primary production in the North Sea over 25 years, associated with reductions in zooplankton abundance and fish stock recruitment. *Global Change Biology*, 24(1):e352–e364.

Carbognin, L., Tosi, L., and Teatini, P. (1995). Analysis of actual land subsidence in Venice and its hinterland (Italy). *Land Subsidence*, pages 129–137.

Chen, Y. D., Carsel, R. F., McCutcheon, S. C., and Nutter, W. L. (1998). Stream temperature simulation of forested riparian areas: I. watershed-scale model development. *Journal of Environmental Engineering*, 124(4):304–315.

Chignoli, C. and Rabagliati, R. (1973). A two-dimensional model for the Lagoon of Venice. pages 203–212.

De Silva, S. S. and Soto, D. (2009). Climate change and aquaculture: potential impacts, adaptation and mitigation. *Climate change implications for fisheries and aquaculture: overview of current scientific knowledge. FAO Fisheries and Aquaculture Technical Paper*, 530:151–212.

Dejak, C., Franco, D., Pastres, R., Pecenik, G., and Solidoro, C. (1992). Thermal exchanges at air-water interfacies and reproduction of temperature vertical profiles in water columns. *Journal of marine systems*, 3(6):465–476.

Di Silvio, G. and D'Alpaos, L. (1972). Il comportamento della Laguna di Venezia esaminato col metodo propagatorio unidimensionale. *Commissione di studio dei provvedimenti per 1a conservazione e difesa della laguna e della citta'di Venezia. Istituto Veneto di Scienze, Lettere ed Arti*, (5).

Dodge, Y. and Commenges, D. (2006). *The Oxford dictionary of statistical terms*. Oxford University Press on Demand.

Dube, A. and Jayaraman, G. (2008). Mathematical modelling of the seasonal variability of plankton in a shallow lagoon. *Nonlinear Analysis: Theory, Methods & Applications*, 69(3):850–865.

Edgar, G. J., Ward, T. J., and Stuart-Smith, R. D. (2018). Rapid declines across Australian fishery stocks indicate global sustainability targets will not be achieved without an expanded network of 'no-fishing'reserves. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 28(6):1337–1350.

European Commission. Directorate-General for Maritime Affairs and Fisheries (2012). *Blue Growth: Opportunities for marine and maritime sustainable growth: Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions*. Publications Office of the European Union.

Ferrarin, C., Maicu, F., and Umgiesser, G. (2017). The effect of lagoons on Adriatic Sea tidal dynamics. *Ocean Modelling*, 119:57–71.

Ferrarin, C. and Umgiesser, G. (2005). Hydrodynamic modeling of a coastal lagoon: the Cabras lagoon in Sardinia, Italy. *Ecological Modelling*, 188(2-4):340–357.

Fredheim, A. and Langan, R. (2009). Advances in technology for off-shore and open ocean finfish aquaculture. In *New Technologies in Aquaculture*, pages 914–944. Elsevier.

Ghezzo, M., Pellizzato, M., De Pascalis, F., Silvestri, S., and Umgiesser, G. (2018). Natural resources and climate change: A study of the potential impact on Manila clam in the Venice lagoon. *Science of The Total Environment*, 645:419–430.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Di, C., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C., and Reiss, P. T. (2020). *refund: Regression with Functional Data*. R package version 0.1-23.

Gunatilaka, A., Moscetta, P., Sanfilippo, L., Savino, E., Dell'Olivo, C., Scardia, F., Gurato, A., and Cisneros-Aguirre, J. (2009). Observations on continuous nutrient monitoring in Venice Lagoon. In *OCEANS 2009*, pages 1–7. IEEE.

Hilborn, R., Amoroso, R. O., Anderson, C. M., Baum, J. K., Branch, T. A., Costello, C., De Moor, C. L., Faraj, A., Hively, D., Jensen, O. P., et al. (2020). Effective fisheries management instrumental in improving fish stock status. *Proceedings of the National Academy of Sciences*, 117(4):2218–2224.

Hussain, M. G., Failler, P., Karim, A. A., and Alam, M. K. (2018). Major opportunities of blue economy development in Bangladesh. *Journal of the Indian Ocean Region*, 14(1):88–99.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F. (2020). *forecast: Forecasting functions for time series and linear models*. R package version 8.12.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2015). Penalized function-on-function regression. *Computational Statistics*, 30(2):539–568.

Jie, C., Jing-Zhang, C., Man-Zhi, T., and Zi-tong, G. (2002). Soil degradation: a global problem endangering sustainable development. *Journal of Geographical Sciences*, 12(2):243–252.

Kokoszka, P. (2012). Dependent Functional Data. *International Scholarly Research Notices*, 2012.

Kokoszka, P. and Reimherr, M. (2017). *Introduction to Functional Data Analysis*. CRC Press.

Liu, Y., Wu, J., Liu, Y., Hu, B. X., Hao, Y., Huo, X., Fan, Y., Yeh, T. J., and Wang, Z.-L. (2015). Analyzing effects of climate change on streamflow in a glacier mountain catchment using an ARMA model. *Quaternary International*, 358:137–145.

Lopes, V. A. R., Fan, F. M., Pontes, P. R. M., Siqueira, V. A., Collischonn, W., and da Motta Marques, D. (2018). A first integrated modelling of a river-lagoon large-scale hydrological system for forecasting purposes. *Journal of Hydrology*, 565:177–196.

Maicu, F., De Pascalis, F., Ferrarin, C., and Umgiesser, G. (2018). Hydrodynamics of the Po River-Delta-Sea System. *Journal of Geophysical Research: Oceans*, 123(9):6349–6372.

Mestekemper, T., Windmann, M., and Kauermann, G. (2010). Functional hourly forecasting of water temperature. *International Journal of Forecasting*, 26(4):684–699.

Mohseni, O., Stefan, H. G., and Erickson, T. R. (1998). A nonlinear regression model for weekly stream temperatures. *Water Resources Research*, 34(10):2685–2692.

Myung, I. J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1):79–95.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roser, M., Ritchie, H., and Ortiz-Ospina, E. (2013). World population growth. *Our World in Data*.

Russo, S., Sillmann, J., and Fischer, E. M. (2015). Top ten European heatwaves since 1950 and their occurrence in the coming decades. *Environmental Research Letters*, 10(12):124003.

Said, S. E. and Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71(3):599–607.

Sguazzero, P., Chignoli, C., Rabagliati, R., and Volpi, G. (1978). Hydrodynamic Numeric Modeling of the Lagoon of Venice. *IBM Journal of Research and Development*, 22(5):472–480.

Shumway, R. H. and Stoffer, D. S. (2000). *Time Series Analysis and Its Applications With R Examples*, volume 3. Springer.

Smith-Godfrey, S. (2016). Defining the blue economy. *Maritime Affairs: Journal of the National Maritime Foundation of India*, 12(1):58–64.

Tacon, A. G. (2020). Trends in global aquaculture and aquafeed production: 2000–2017. *Reviews in Fisheries Science & Aquaculture*, 28(1):43–56.

Tipler, P. A. (1999). *Physics for Scientists and Engineers: Regular Version, Ch. 1-35 and 39*. MacMillan.

Ullah, S. and Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, 13(1):43.

Umgiesser, G. (1997). Modelling the Venice lagoon. *International Journal of Salt Lake Research*, 6(2):175–199.

Umgiesser, G., Canu, D. M., Cucco, A., and Solidoro, C. (2004). A finite element model for the Venice Lagoon. Development, set up, calibration and validation. *Journal of Marine Systems*, 51(1-4):123–145.

Umgiesser, G., Ferrarin, C., Cucco, A., De Pascalis, F., Bellafiore, D., Ghezzo, M., and Bajo, M. (2014). Comparative hydrodynamics of 10 Mediterranean lagoons by means of numerical modeling. *Journal of Geophysical Research: Oceans*, 119(4):2212–2226.

UN Department for Economic and Social Affairs (2019). World population prospects 2019: Highlights. *New York (US): United Nations Department for Economic and Social Affairs*.

UN General Assembly (2015). Transforming our world: the 2030 Agenda for Sustainable Development. *Division for Sustainable Development Goals: New York, NY, USA*.

Vasilakopoulos, P., Maravelias, C. D., and Tserpes, G. (2014). The alarming decline of Mediterranean fish stocks. *Current Biology*, 24(14):1643–1648.

Vaz, N., Dias, J. M., Leitao, P., and Martins, I. (2005). Horizontal patterns of water temperature and salinity in an estuarine tidal channel: Ria de Aveiro. *Ocean Dynamics*, 55(5-6):416–429.

Wit, E., Heuvel, E. v. d., and Romeijn, J.-W. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236.