Università
Ca'Foscari
Venezia

Master Degree in
Data Analytics For Business and Society

Predicting the Venice tide using State Space models and the Kalman filter

Graduand:
Giorgio Bernacchi Monti

Matriculation Number:
886263

Supervisor:
Prof. Dario Palumbo

Academic Year:
2024-25

Abstract

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The city of Venice is located inside the Venice Lagoon, a semi-enclosed basin in the northern Adriatic Sea, connected to the sea through three main inlets. Due to its low elevation above mean sea level, the city is frequently affected by storm surge events known as "Acqua Alta". Over the past century, the frequency of flooding events has increased due to local land subsidence and climate change.
In October 2018, a storm surge reached a water level of 156 cm. In November 2019, on the 12th and 13th, water levels reached 189 cm, one of the highest levels ever recorded. Such events are considered exceptional, as only 25 have been recorded in the last 150 years.
These events pose a serious threat to the artistic, cultural, and economic environment of the city. In response, significant investments have been made in safeguarding Venice, through the construction of flood barriers (MOSE), and research initiatives aimed at studying tide levels and developing forecasting systems to predict water levels within the lagoon.
The objective of this work is to evaluate the impact of each tidal component on the overall sea level and to provide forecasts for horizons ranging from 24 to 168 hours. This is achieved using a State Space Model (SSM) combined with the Kalman Filter for recursive estimation and forecasting. The advantage of this approach over traditional prediction techniques is the creation of a dynamic system that continuously adapts its parameters as new observations are incorporated. This allows the model to account for sudden environmental shocks.
The thesis is structured as follows: Chapter 2 establishes the mathematical foundation for model definition and signal decomposition. It is structured in several sections, beginning with the Wold Decomposition Theorem, which explains how any stochastic process can be represented as a linear combination of its singular and regular components. The second section introduces the state space model for different processes, discussing the main advantages of the application of various rep-

resentations. The final section describes the Kalman filter and the methodology used for prediction and forecasting.
Chapter 3 discusses data sourcing, variable transformation, and specification. The transformations applied to each variable were designed to reflect the physical effects on water levels, specifically through the use of wind stress, inverse barometric pressure, and pressure gradients.
Chapter 4 presents the empirical results based on 2019 data. Two main approaches are compared: the first treats the tidal cycle as deterministic, adding cyclical stochastic components to capture any unexplained effects. The second treats the eight primary astronomical harmonics as stochastic terms, allowing their amplitude and phase to vary over time.

1.1. Literature review

The theoretical foundation for tidal modeling is the Wold Decomposition Theorem (Wold, 1938), which states that any discrete stationary process can be represented as the sum of a singular (deterministic) and a regular (stochastic) component. As discussed in Sartore (1975), the Venice tide can be decomposed into a deterministic astronomical tide and an unpredictable stochastic surge. The astronomical components are governed by the gravitational interactions between the Earth, Moon, and Sun (Doodson, 1921), while the stochastic residue captures the fluctuations driven by meteorological forcing.
Beyond cyclical patterns, Venice is characterized by long-term geological trends that modify the sea-level reference. Sartore (1975) identified a local land subsidence rate of 2.479 mm per year. Regarding short-term variations, Vieira et al. (1993) identified the main exogenous drivers of the stochastic surge: wind—specifically the Sirocco (SE) and Bora (NE)—atmospheric pressure, and pressure gradients. Their study also demonstrated that statistical regression models can achieve a Root Mean Square Error (RMSE) comparable to complex hydrodynamic simulations, particularly for operational short-term periods.
During the early 2000s, the focus shifted from simple linear regression toward more complex expert systems. As documented by Tosoni and Canestrelli (2011), the Centro Previsioni e Segnalazioni Maree (CPSM) developed a family of ARMAX (Auto-Regressive Moving Average with Exogenous inputs) models, stratified by classes of seasonality and atmospheric patterns to account for the non-linear response of the lagoon.
However, these systems relied on fixed coefficients based on each class. Their limitation became evident during the exceptional floods of December 2008, when the static parameters failed to adapt to the shock of the meteorological conditions. This highlighted the need for a system that could update its parameters in a dynamic way.

To overcome these limitations, recent research has shifted toward more flexible, adaptive approaches. A key contribution is Bajo (2010), who applied the Kalman Filter as a post-processing model. Unlike the earlier models described by Tosoni (2011), which rely on a fixed set of rules for different seasons, the Kalman Filter operates as a continuous, self-updating estimation procedure.

The strength of this approach lies in its error-handling mechanism. At each time step, the model compares its prediction to the observed water level from tide gauges. The difference — known as the innovation — is used to recalibrate the model in real time. This self-correcting ability allows the system to absorb sudden environmental shocks, such as rapid changes in wind or pressure, even those without historical precedent.

The importance of this flexibility was clearly demonstrated during the historic flood of November 2019, when water levels reached a devastating 189 cm. As noted by Umgiesser et al. (2021), the increasing unpredictability of the Adriatic climate means that traditional rigid models are no longer sufficient. Forecasting systems must therefore be capable of adapting in real time, rather than relying solely on historical storm patterns.

Building on Sartore (1975), this research proposes a different treatment of the singular component, the cyclical pattern, within the Wold decomposition. While traditional models treat the eight principal harmonics as fixed deterministic constants, this work implements a State Space Model in which these harmonics are modeled as stochastic variables. By allowing amplitude and phase to evolve dynamically through the Kalman Filter, the model captures variability that would otherwise remain trapped in the error term of traditional fixed-coefficient approaches such as ARMAX.

Chapter 2

Theoretical and Mathematical framework

2.1. Wold Decomposition Theorem

According to Wold's decomposition theorem (Wold, 1938), any discrete stationary process with finite variance can be represented as a two-component stationary process. In this framework, the random variable y_t is composed of a singular process z_t and a regular process x_t, characterized by the following properties:

A) $y_t = z_t + x_t$,

B) z_t and x_t are uncorrelated,

C) z_t is a singular process,

D) ε_t is a non-autocorrelated process,

E) $x_t = \sum_{j=1}^{\infty} \beta_j \varepsilon_{t-j}$ where the coefficients represent real finite numbers, $\sum_{j=0}^{\infty} \beta_j^2 < \infty$.

The fundamental requirement of the theorem is that the process y_t possesses a stationary covariance. This condition is met when the individual components of the random process are time-invariant. Wold defines a process as stationary when the random variable y_t satisfies the following criteria:

A) For any arbitrary subset of t, the joint distribution of the random variable y_t is well-defined.

B) Given that, the distribution function and the probability functions of y_t are denoted by F and P, and their relationship is defined as follows:

$$F(t_1, ..., t_n; \mu_1, ..., \mu_n) = P[y(t_1) < \mu_1, y(t_2) < \mu_2, ..., y(t_n) < \mu_n] \quad (2.30)$$

Given a subset of t and any permutation $i_1, ..., i_n$, the distribution function must satisfy:

$$F(t_{i_1}, t_{i_2}, ..., t_{i_n}; \mu_{i_1}, \mu_{i_2}, ..., \mu_{i_n}) = F(t_1, ..., t_n; \mu_1, ..., \mu_n) \quad (2.31)$$
$$F(t_1, ..., t_m; \mu_1, \mu_m) = F(t_1, ..., t_n; \mu_1, ..., \mu_m, ...\infty) \quad (2.32)$$

These conditions indicate that a process is stationary if its joint distribution F is invariant to time translations, meaning that the probability of y_t falling below any threshold μ remains constant over time. As a result, y_t maintains a constant mean and finite variance throughout the series.
While Wold's theorem provides a universal framework for stationary processes, its application to tidal modeling requires a precise specification of each component. Sartore (1975) characterizes the Venice tide as a signal in which the singular and regular components represent distinct environmental forces.
The singular component, z_t, represents the astronomical tide, modeled as a periodic and deterministic process. It is defined by the following equation:

$$y_t = \alpha \cos \lambda t + \beta \sin \lambda t$$

where α and β are fixed parameters and $\lambda \in [0, \pi]$. In this form, the process is non-stationary since its value depends directly on time. To ensure stationarity, α and β are instead treated as random variables with zero mean and constant variance, denoted σ_α^2 and σ_β^2 (Harvey, 1989). Under these conditions, the process maintains a mean of zero, a constant variance, and a covariance that depends on the lag τ rather than absolute time:

$$E(y_t) = 0 \quad (2.33\text{ a})$$
$$Var(y_t) = \sigma^2 \quad (2.33\text{ b})$$
$$Cov[y_t, y_{t-\tau}] = \sigma^2 \cos \lambda\tau \quad (2.33\text{ c})$$

These results confirm that stationarity of the cycle requires treating α and β as random variables rather than fixed parameters. This is verified by the fact that the mean is zero, the variance is constant, and the covariance depends solely on the lag shift.
The atmospheric component v_t comprises a deterministic trend and a stochastic process driven by atmospheric forces, representing the regular components of the signal. The structure of v_t is defined as follows:

$$\begin{cases} v_t = d_t + s_t \\ s_t = \sum_{j=1}^{k} \beta_j \omega_{jt} + x_t \\ d_t = \alpha_0 + \alpha_1 t \\ x_t = \phi x_{t-1} + \varepsilon_t \end{cases} \quad (2.34)$$

where

- Deterministic trend (d_t): This variable defines the non-stochastic evolution of the sea level. Modeled as a linear trend, the coefficient α_1 represents the combined effects of subsidence (the downward vertical movement of Venice) and eustatism (the rise of the Adriatic sea level).
- Stochastic atmospheric component (s_t): This captures the influence of atmospheric factors, such as local wind speed, wind direction, and barometric pressure along the Adriatic coast. The term x_t represents the deviation not explained by the atmospheric forces, modeled as a first-order autoregressive process ($AR(1)$). To ensure the stationarity of the process, the parameter ϕ is constrained such that $|\phi| < 1$.

Together, these elements allow the model to partition the total sea level into a predictable geological drift and a mean-reverting stochastic process, fulfilling the requirements of the regular component in the Wold decomposition.

2.2. State Space model

The State Space Model (SSM) is a versatile framework for analyzing both univariate and multivariate time series. Once expressed in state-space form, the system can be estimated using a recursive algorithm known as the Kalman Filter, which performs filtering and smoothing through prediction error decomposition.
The general Gaussian representation of a state-space model is defined by the following system of equations:
The measurement equation (2.1) describes the relationship between the observed variable y_t and the unobserved (hidden) states α_t:

$$y_t = Z_t\alpha_t + d_t + \varepsilon_t, \quad t = (1, ..., n) \quad (2.1)$$

$$E(\varepsilon_t) = 0 \quad Var(\varepsilon_t) = H_t \quad (2.2)$$

In this context, y_t is an $N \times 1$ vector representing the observed series at time t. The state vector α_t is an $m \times 1$ vector of unknown parameters. Furthermore, Z_t is an $N \times m$ design matrix, d_t is an $N \times 1$ vector of deterministic components, and ε_t is an $N \times 1$ vector of disturbances (white noise) with mean zero and covariance matrix H_t.
The unobservable elements of α_t evolve according to the transition equation (2.3).

$$\alpha_t = T\alpha_{t-1} + c_t + R_t\eta_t \quad \text{t = (1, ..., n)} \quad (2.3)$$

This equation describes a first-order Markov process, meaning that the current state depends only on the immediately preceding state. Here, T_t is an $m \times m$ transition matrix that determines how the state evolves over time; c_t is an $m \times 1$ vector of constant terms; R_t is an $m \times g$ selection matrix that maps the disturbances to their respective states; and η_t is a $g \times 1$ vector of uncorrelated disturbances with mean zero and covariance matrix Q_t.

$$E(\eta_t) = 0 \quad Var(\eta_t) = Q_t \qquad (2.4)$$

To ensure the model is well-defined, the following assumptions are required:

1. The initial state α_0 has a mean a_0 and a covariance matrix P_0

$$E(\alpha_0) = a_0 \quad Var(\alpha_0) = P_0 \quad (2.5)$$

2. The disturbance terms ε_t and η_t are uncorrelated with each other for all time periods and uncorrelated with the initial state α_0

$$E(\varepsilon_t \eta_s') = 0 \quad \text{for all s,t} = 1, ..., T \quad (2.6)$$
$$E(\varepsilon_t \alpha_0) = 0 \quad E(\eta_t \alpha_0) = 0 \text{ for all t = 1, ..., T} \qquad (2.7)$$

2.2.1. Specification of structural components in SSM

This section derives the State Space Model (SSM) for various stochastic processes. By partitioning the state vector α_t, it is possible to represent distinct time series components - such as trends, cycles, and seasonal patterns - within the unified framework established in equations (2.1) and (2.3).
The trend can be formally described using the $ARIMA(p, d, q)$ specification:

$$(1 - \textstyle\sum_{i=1}^{p} \phi_i L)(1 - L)^d y_t = (1 + \sum_{i=1}^{q} \theta_q L)\varepsilon_t \qquad (2.8)$$

Depending on the order of the process, different trend effects can be isolated. The following sections outline the composition of the state vector for several cases, highlighting the flexibility of the SSM in allowing multiple representations of α_t.
Autoregressive process
An autoregressive process of any order can be cast in state space form starting from the AR(p) equation.

$$(1 - \textstyle\sum_{i=1}^{p} \phi_p) y_t = \varepsilon_t$$

Following Harvey (1989), α_t is defined to contain the current and lagged values of the process:

$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-p} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} \phi_1 & \phi_2 & \vdots & \phi_p \\ 1 & 0 & \vdots & 0 \\ 0 & 1 & \vdots & 0 \\ 0 & 0 & \ddots & 0 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p-1} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{R_t} \eta_t \qquad (2.9)$$

$$Z = \begin{bmatrix} 1 & 0 & \vdots & 0 \end{bmatrix} \qquad Q = RR'\sigma^2$$

In this representation, T acts as a companion matrix. Its first row contains the autoregressive parameters, while an identity matrix of dimension $(p-1) \times (p-1)$ is placed below it, to ensure the correct temporal shift between lags.
Moving average processes of order q are described by the following equation:

$$y_t = (1 + \textstyle\sum_{i=1}^{q} \theta_i)\varepsilon_t \quad 2.10$$

$$\underbrace{\begin{bmatrix} y_t \\ \theta_1\varepsilon_t \\ \theta_2\varepsilon_{t-1} \\ \vdots \\ \theta_q\varepsilon_{t-q} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_{t-1} \\ \theta_1\varepsilon_{t-1} \\ \theta_2\varepsilon_{t-2} \\ \vdots \\ \theta_q\varepsilon_{t-q} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} 1 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_q \end{bmatrix}}_{R_t} \eta_t \quad (2.11)$$

$$Z = \begin{bmatrix} 1 & 0 & \vdots & 0 \end{bmatrix} \qquad Q = RR'\sigma^2$$

Moving average process
For any MA process, the state vector is built in order to keep track of the disturbance. The matrix T will contain zeros in the first column and an identity block of dimension $(q-1) \times (q-1)$ to correctly move the shocks. The R matrix is crucial because it determines how each disturbance propagates through the levels of the state vector.
Differencing is applied when a time series is non-stationary, typically when the roots of the characteristic equation lie on or outside the unit circle (Box & Jenkins, 1994). Box & Jenkins show how the process can be made stationary by differencing, expressing the current value of y_t in terms of previous values of y's and current and previous values of ε's.

$$(1-L)^d y_t = \varepsilon_t$$

In the Box & Jenkins SSF, the state vector is defined by current and lagged observations:

$$\delta_k = (-1)^k(\tbinom{d}{k}) \qquad \tbinom{d}{k} = \tfrac{d!}{k!(d-k)!} \qquad \delta_k^* = \textstyle\sum_{k=0}^{i} -\delta_k$$

$$\underbrace{\begin{bmatrix} y_t \\ \vdots \\ y_{t-d} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} \delta_1 & \vdots & \delta_d \\ 1 & 0 & 0 \\ 0 & \ddots & 0 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_{t-1} \\ \vdots \\ y_{t-d-1} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{R_t} \eta_t \quad (2.12)$$

Equation (2.12) is generalized for differencing of any order d. The first row contains coefficients obtained from a binomial expansion, capturing the dependence of the current state and the previous levels. The remaining rows form an identity matrix of dimension $(d-1) \times (d-1)$, ensuring the correct time shift between states.
However, Durbin and Koopman (2001) argue that this approach - where hidden states are simply lagged observations - is less informative than the structural form.

Structural time series models explicitly decompose the series into unobserved components such as level (μ_t), slope (β_t), and acceleration (γ_t). As Harvey (1989) notes, these models are attractive because they allow for the direct estimation of these latent environmental or economic forces. For a third-order integrated process $I(3)$, the structural SSF is:

$$\underbrace{\begin{bmatrix}\mu_t\\ \beta_t\\ \gamma_t\end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix}1 & 1 & 0\\ 0 & 1 & 1\\ 0 & 0 & 1\end{bmatrix}}_{T}\underbrace{\begin{bmatrix}\mu_{t-1}\\ \beta_{t-1}\\ \gamma_{t-1}\end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}}_{R_t}\underbrace{\begin{bmatrix}\eta_t\\ \zeta_t\\ \omega_t\end{bmatrix}}_{\eta_t} \tag{2.13}$$

Ex. I(3) with level, slope and acceleration

While the Box-Jenkins form (2.12) is computationally simpler, the structural form (2.13) provides a deeper physical interpretation of hidden states, which is particularly useful in tidal modeling for separating geological drift and stochastic weather effects.

ARMA and ARIMA

ARMA processes combine AR(p) and MA(q) within the transition matrix T and the selection matrix R, and are used to analyze stationary series. Following Hamilton (1994, pp. 374–375), the state vector contains the lagged levels of the series up to:

$$m = max(p, q + 1)$$

$$\underbrace{\begin{bmatrix}y_t\\ y_{t-1}\\ \vdots\\ y_{t-m+1}\end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix}\phi_1 & \phi_2 & \vdots & \phi_m\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & \ddots & 0\\ 0 & 0 & 0 & 1\end{bmatrix}}_{T}\underbrace{\begin{bmatrix}y_{t-1}\\ \vdots\\ y_{t-m}\end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix}1\\ \theta_1\\ \vdots\\ \theta_q\end{bmatrix}}_{R_t}\eta_t \tag{2.14}$$

Matrix (2.14) illustrates the configuration of the transition matrix. The first row shows the autoregressive component, defining the stochastic dependence of the observations. Below it, an identity matrix of dimension $(m-1)\times(m-1)$ controls the correct time shift of the observations. The moving average component is entirely governed by the matrix R, which extends the shocks across the different levels of the state vector.

For an ARIMA model, a key distinction arises based on the Box & Jenkins approach and the structural form provided by Harvey (1989). With Box & Jenkins, the hidden state can be represented as a combination of the different levels. In the case of ARIMA, the differencing operator can be incorporated into the transition matrix using binomial coefficients, as shown in (2.15).

$$\delta_k = (-1)^k \binom{d}{k} \quad \binom{d}{k} = \frac{d!}{k!(d-k)!}$$
$$\phi_j^* = \sum_{k=0}^{j} -\delta_k \phi_{j-k}$$
$$m = max(p+d, q+1)$$
$$\underbrace{\begin{bmatrix} y_t \\ y_{t-1} \\ \vdots \\ y_{t-m+1} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} \phi_1^* & \phi_2^* & \vdots & \phi_{m-1}^* & \phi_m^* \\ 1 & 0 & \vdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \vdots & 1 & 0 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-m} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} 1 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}}_{R_t} \eta_t \quad (2.15)$$

As noted above, Harvey & Durbin highlight the advantage of using structural forms, which incorporate differencing directly into the hidden state to study each specific unobserved component of the process (2.16).

$$\underbrace{\begin{bmatrix} y_t \\ \Delta y_t \\ \Delta^2 y_t \\ \Delta^2 y_{t-1} \\ ma_{1,t} \\ ma_{2,t} \\ ma_{3,t} \end{bmatrix}}_{\alpha_t} = \underbrace{\begin{bmatrix} 1 & 1 & \phi_1 & \phi_2 & 1 & 0 & 0 \\ 0 & 1 & \phi_1 & \phi_2 & 1 & 0 & 0 \\ 0 & 0 & \phi_1 & \phi_2 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} y_{t-1} \\ \Delta y_{t-1} \\ \Delta^2 y_{t-1} \\ \Delta^2 y_{t-2} \\ ma_{1,t-1} \\ ma_{2,t-2} \\ ma_{3,t-3} \end{bmatrix}}_{\alpha_{t-1}} + \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}}_{R_t} \eta_t \quad (2.16)$$

ARIMA(2, 2, 3) using structural approach

where the MA(q) process is defined inside the hidden state as a Markov process (2.17).

$$\begin{aligned} ma_{3,t} &= \theta_3 \varepsilon_t \\ ma_{2,t} &= ma_{3,t-1} + \theta_2 \varepsilon_t = \theta_3 \varepsilon_{t-1} + \theta_2 \varepsilon_t \\ ma_{1,t} &= ma_{2,t-1} + \theta_1 \varepsilon_t = \theta_3 \varepsilon_{t-2} + \theta_2 \varepsilon_{t-1} + \theta_1 \varepsilon_t \end{aligned} \quad (2.17)$$

The structural form produces a state vector with multiple components, as shown in (2.17). The diagonal blocks of the transition matrix correspond to the main components - I(2), AR(2), MA(3) - while the off-diagonal blocks capture their correlation (2.18).

$$\begin{bmatrix} \text{I(d)} & \text{Corr I(d) and AR(p)} & \text{Corr I(d) MA(q)} \\ \text{Corr AR(p) I(d)} & \text{AR(P)} & \text{Corr AR(p) MA(q)} \\ \text{Corr MA(q) I(d)} & \text{Corr MA(q) AR(p)} & MA(q) \end{bmatrix} \quad (2.18)$$

This form enables a precise distinction of the hidden state components, at the cost of higher computational time due to the increase of all matrix dimensions.

Cycle

Any deterministic cycle can be represented as:

$$y = cos(x) \ (2.19)$$

where x is an angle measured in radians. Since a full cycle spans 2π radians, y covers its full range as x moves from 0 to 2π. This pattern is periodic and repeats for any integer k, as shown in (2.20).

$$cos(x + 2k\pi) = cos(x) \quad (2.20)$$

The movement of the sine function follows a similar periodic pattern, and to express the variable y as a cyclical function of time, we introduce the parameter λ, measured in radians, known as angular frequency.

The variable x in equation (2.19) is substituted with the variables λt. By assigning a different value to λ, the function can expand or contract along the time axis. The cycle period is defined as the time for y to complete a full oscillation, equal to $\frac{2\pi}{\lambda}$.

The equation can be further generalized by introducing the amplitude ρ and the phase θ, measured in radians (2.21).

$$y = \rho\cos(\lambda t - \theta) = \rho\cos\lambda(t - \zeta), \zeta = \theta/\lambda \quad (2.21)$$

where ζ represents the time shift. Since the explicit phase term introduces non-linearity in estimation, it is preferable to express the cycle as a linear combination of cosine and sine functions:

$$y = \alpha\cos(\lambda_t) + \beta\sin(\lambda_t) \quad (2.22)$$

where α and β determine the amplitude and phase of the oscillation. The relation between this representation and the one in (2.21) is the following:

$$\rho^2 = \alpha^2 + \beta^2, \quad \theta = \tan^{-1}(\tfrac{\beta}{\alpha})$$

To extend this representation to a time series, y_t is defined as a linear function of n trigonometric terms (Harvey, 1989).

$$n = \begin{cases} T/2 \text{ if } T \text{ is even} \\ (T-1)/2 \text{ if } T \text{ is odd} \end{cases} \quad (2.22)$$

$$\lambda_j = 2\pi j/T \quad (2.23)$$

The parameter n in (2.22) defines the number of harmonics in the series. It is evaluated as half of the total number of observations, based on the Nyquist number, $\lambda = \pi$. This limit defines the highest frequency that can be captured, indicating that at least two data points are required to define a single cycle.

$$y_t = T^{-\frac{1}{2}}\alpha_0 + \sqrt{2/T}\sum_{j=1}^{n-1}(\alpha_j\cos\lambda_j t + b_j\sin\lambda_j t) + T^{-\frac{1}{2}}\alpha_n(-1)^t \ (2.24)$$

Equation (2.24) shows how each observation y_t can be decomposed into distinct, orthogonal, and uncorrelated components. This can be explained by analyzing the three main components of the equation. The first term α_0 defines the base level around which the harmonic components oscillate:

$$\alpha_0 = T^{-\frac{1}{2}} \sum_{t=1}^{T} y_t \underbrace{\cos 2\pi 0/T}_{1} = T^{-\frac{1}{2}} \sum_{t=1}^{T} y_t$$

α_n corresponds to the highest frequency of the series, where the angular frequency is equal to the Nyquist number, $\lambda = \pi$.

$$\alpha_n = T^{-\frac{1}{2}} \sum_{t=1}^{T} y_t \underbrace{\cos(2\pi n/T)}_{-1} = T^{-\frac{1}{2}} \sum_{t=1}^{T} y_t(-1)^t \text{ for n} = \max(\text{n})$$

The central term decomposes the complex signal into its constituent harmonics. Each index j represents a unique frequency λ_j, allowing the model to capture multiple cyclical components simultaneously. Because these trigonometric terms are defined at frequencies $\lambda_j = 2\pi j/T$, they form an orthogonal system. This ensures that each harmonic is mathematically independent, preventing correlation between the terms and allowing for a clean separation of different effects. The terms β_0 and β_n are not included in the system because $\sin(0)$ and $\sin(\pi)$ are equal to 0.
The Fourier representation in equation (2.24) demonstrates that a signal can be decomposed into a set of fixed deterministic sine and cosine waves. However, for many time series the use of a deterministic model is inappropriate, because the amplitude and phase of these cycles can change over time. The transition from a deterministic cycle to a stochastic one is achieved by allowing the coefficients to evolve dynamically within the SSF. The central term of equation (2.24) is represented as a recursive state-space system for each frequency λ_c.

$$\underbrace{\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix}}_{\alpha_t} = \rho \underbrace{\begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix}}_{\alpha_{t-1}} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{2.25}$$

In this formulation, the components of the *state vector* (ψ_t, ψ_t^*) function as coefficients (α_j and β_j). The matrix T defines the deterministic part of the cycle by shifting the phase of the components by the angular frequency λ_c, at each time step of the series. The damping factor ρ controls the persistence of the cycle. For $0 < \rho < 1$, the cycle is stationary, with its amplitude decaying gradually over time. The disturbance terms (κ_t, κ_t^*) are uncorrelated; these shocks allow the amplitude and phase of the cycle to evolve over time. The T matrix can accommodate multiple uncorrelated cycles by stacking the individual rotation matrices block-diagonally.

2.3. Kalman filter

The Kalman filter is an algorithm named after R.E. Kalman (1960, 1963), for sequentially updating a linear projection of a system. In this case, the state-space representation discussed in the previous subsection is used to calculate the optimal estimator of the state vector given all the currently available information.
Once the end of the series is reached, optimal predictions of future observations can be made. Additionally, a smoothing estimation based on backward recursion can be performed to obtain optimal estimates of the state vector at all time points.
Let a_t denote the optimal estimator of the state vector α_t, based on all observations including y_t. Its mean square error (MSE) is defined as P_t and can be evaluated as:

$$P_t = E[(\alpha_t - a_t)(\alpha_t - a_t)'] \quad (2.26)$$

If we are at time t-1, and the values of a_{t-1} and P_{t-1} are given, the optimal estimator of the vector α_t is given by the prediction equation:

$$a_{t|t-1} = T_t a_{t-1} + c_t$$
$$P_{t|t-1} = T_t P_{t-1} T_t' + R_t Q_t R_t, \quad \text{t = 1, ..., T} \quad (2.27)$$

The MSE matrix P_t, represents the covariance of the estimated errors, and gives a measure of the model confidence. High diagonal elements of P_t indicate high uncertainty in the corresponding hidden state, while low values indicate convergence of the filter. The predicted value of y_t is:

$$\tilde{y}_{t|t-1} = Z_t a_{t|t-1} + d_t$$

The MSE error evaluated from the prediction error is given by the innovation vector:

$$v_t = y_t - \tilde{y}_{t|t-1} = Z_t(\alpha_t - a_{t|t-1}) + \varepsilon_t, \quad t = 1, ..., T \quad (2.28\text{ a})$$
$$F_t = Z_t P_{t|t-1} Z_t' + H_t \quad (2.28\text{ b})$$

The matrix F_t, called innovation variance, captures the uncertainty of the prediction by combining the state uncertainty with the measurement noise H_t. Once new observations become available, the estimations of the state can be made using the updating equation:

$$a_t = a_{t|t-1} + P_{t|t-1} Z_t' F_t^{-1}(y_t - Z_t a_{t|t-1} - d_t) \quad (2.29\text{ a})$$
$$P_t = P_{t|t-1} - P_{t|t-1} Z_t' F_t^{-1} Z_t P_{t|t-1} \quad (2.29\text{b})$$

The main part of the equation is given by v_t, the innovation, because it defines the difference between observations and predicted values. A larger deviation of this value indicates a greater change in the states.
Equations (2.29a) and (2.29b) define how the state is updated when a new observation is incorporated into the model. The term $P_{t|t-1} Z_t' F_t^{-1}$ in equation (2.29) acts as a weight, controlling the influence of the innovation on the state update.
Once the filter has processed all T observations, it can be used to make predictions

based on the last observed state vector. The optimal estimator for any future period $T + l$ is obtained by projecting the final result of the filter through the transition matrix T.

$$a_{T+l|T} = T_{T+l} a_{T+l-1} + c_{T+l} \quad l = 1, 2, ...$$

During the forecast phase, the matrix $P_{T+l|T}$ quantifies the reliability of the prediction. Unlike the filtering phase, where new observations reduce the error covariance via the innovation equation (2.28), the forecast accumulates uncertainty at each step due to the process noise Q.

$$P_{T+l|T} = T_{T+l} P_{T+l-1} T'_{T+l} + R_{t+l} Q_{t+l} R'_{T+l}, \quad l = 1, 2, ...$$

The systems in a state space model depend on a set of unknown parameters, defined in an $n \times 1$ vector ψ, referred to as hyperparameters. After the filter is initialized, the hyperparameters can be evaluated via Maximum Likelihood estimation. Assuming the disturbances are Gaussian, the log-likelihood function is constructed via prediction error decomposition (Harvey, 1989):

$$\log L(\psi) = -\frac{NT}{2} \log 2\pi - \frac{t}{2} \sum_{t=1}^{T} \log |F_t| - \frac{1}{2} \sum_{t=1}^{T} v'_t F_t^{-1} v_t$$

The Kalman filter recursively evaluates the innovation, v_t and its covariance matrix, F_t, allowing the likelihood to be estimated. A non-linear optimization algorithm is then used to find the parameters that maximize the log-likelihood function.
The initialization of the filter depends on the stationarity of the system. Given the initial state α_0 and its error covariance matrix P_0, if the state is stationary, the filter can be initialized with the mean and the covariance matrix of the unconditional distribution of α_t.

$$\alpha_0 = E[\alpha_t] \quad vec(P_0) = [I - (T \otimes T)]^{-1} vec(Q)$$

When the state vector is non-stationary and prior information is not available, the initialization is done via diffuse prior distribution. A state is defined as diffuse if its covariance matrix, P_0, is arbitrarily large, indicating total uncertainty regarding the initial location of the hidden state.

$$\alpha_0 = 0 \quad P_0 = \kappa I, \kappa \to \infty$$

The value of κ is set to a large value such that the filter weight is placed entirely on the incoming measurement during the first iterations.

Chapter 3

Data and model methodology

3.1. Data Description and Pre-processing

This section translates the mathematical components of the Venice tidal model into observable variables. Following the structural decomposition of Sartore (1975), described in Section 2.1, and the hydrodynamic framework of Vieira (1993), multiple datasets were constructed by aligning high-frequency tidal observations with localized meteorological parameters. The dependent variable y_t represents the sea level in centimeters at the Punta della Salute station, the primary benchmark for the Venice Lagoon and the reference point for all tidal measurements, defined as the Zero Mareografico Punta della Salute (ZMPS).
Data were extracted from the Rete Mareografica Laguna di Venezia (RMN) for the period 1989 - 2024. The year 2019 was selected as the primary reference, as it ensures synchronization with the mean sea level and subsidence estimates from Sartore (1975). The singular component z_t captures the cyclical behaviour of the astronomical tide; each component is characterized by amplitude, phase, and angu-

		M_2	S_2	N_2	K_2	K_1	O_1	P_1	S_1
Ampiezza (cm)	**A**	24.1	13.8	3.9	3.8	17.0	5.0	5.5	1.4
Fase (gradi)	φ	187.5	324.0	264.8	134.9	88.7	308.5	97.7	271.0
Velocità angolare (gradi/ora)	ω	28.9841042	30.0	28.4397295	30.0821373	15.0410686	13.9430356	14.9589314	15.0000020

Figure 3.1: Harmonic constants for the eight primary tidal constituents at Punta della Salute

lar frequency. These values were extracted from the Calendario delle Maree 2019

(Comune di Venezia) for the Punta della Salute station (Figure 3.1). Amplitudes are measured in centimeters, while phases and angular frequencies are given in degrees and degrees per hour. These were converted to radians using the NumPy `deg2rad` function, as required by the transition matrix T. The resulting values initialize the state vector α_t at t=0, providing a physically grounded starting point for the Kalman Filter.

The stochastic component x_t captures the atmospheric contribution to sea level variability. Following Vieira (1993), the main drivers are wind speed and direction measured on-site, and air pressure gradients at multiple points along the Adriatic coast (Figure 3.2). Wind data were sourced from the Rete Mareografica Nazionale

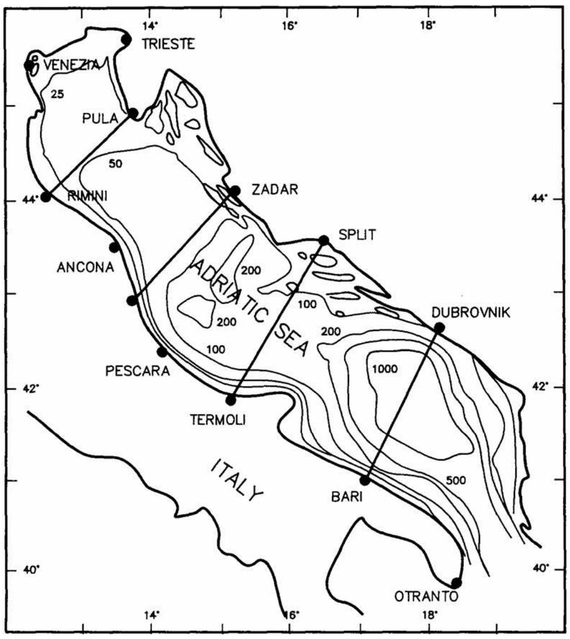

Figure 3.2: Stations for measuring air pressure gradient Vieira 1993

(RMN) at the Diga Sud Lido station, covering the period 2014–2022. This station was selected due to its position at the lagoon edge, where it captures wind effects directly with minimal interference from urban structures.

Wind speed was provided in m/s and direction in degrees North (N°). A key challenge with raw directional data (0°, 360°) is the circular discontinuity at North, where a small physical change can cause a large numerical jump, potentially distorting the Kalman Filter innovation. To ensure numerical stability, directions were remapped to a continuous scale between -1 and 1 using a cosine transformation. To represent the physical relationship between wind and sea level, these measurements are evaluated as wind stress components. This measure defines the transfer of momentum from the atmosphere to the surface of the water. Wind speed and direction are decomposed into two main components, τ_{tx} and τ_{ty}, representing stress along the

East-West and North-South axes. This decomposition allows the state-space model to capture the linear effect of these forces on the tidal signal.
Air pressure data were collected from multiple stations along the Adriatic (Figure 3.2). Italian coastal data were sourced from the RMN, while Croatian coastal data were obtained from the ERA5 reanalysis dataset (Copernicus Climate Change Service, hourly data from 1940 to present). Since ERA5 data are provided in GRIB format, a custom Python script using the `cfgrib` package was used to extract hourly values at the station coordinates identified by Vieira (1993). Latitude and longitude for each station were sourced from the Croatian Meteorological and Hydrological Service (DHMZ).
All pressure observations are in hPa. Two components are derived: the inverse barometric pressure (IBP), which measures the sea level response to local pressure variations, and the pressure gradient, which captures the effect of pressure differences between stations. The IBP at Punta della Salute is computed as the deviation from the series mean. The pressure gradient for all other stations is computed as the pressure difference between Venice and each station, divided by their distance, calculated using the Haversine formula.
The deterministic parameters for the regular component were initialized following the findings of Sartore (1975). The Mean Sea Level (MSL) was set at 31 cm relative to the Zero Mareografico Punta della Salute (ZMPS), providing the baseline intercept α_0 for the measurement equation. The long-term trend, reflecting the combined effects of subsidence and eustatism, was set at 2.479mm/year. To maintain consistency with the hourly frequency of the tidal observations, this value was converted to cm and divided by 8760, yielding a constant slope α_1 expressed in cm/h. This precise calibration allows the model to partition the total sea level into a predictable geological drift and a mean-reverting stochastic surge, fulfilling the theoretical requirements of the Wold Decomposition.

3.1.1. Exploratory Data Analysis

The dependent variable y_t representing the tide levels at Punta della Salute, follows a distribution that approximates a Gaussian curve (Figure 3.3). Data for the year 2019 are centered on a mean of 36.17 cm and a variance of 851.67. The high variance and the heavy tails of the distribution, extending from -75 to 200 cm, are driven by the stochastic surge. Acqua alta events create extreme observations, driven by meteorological shocks such as wind stress, inverse barometric pressure, and pressure gradient. The multiple local peaks near the top of the distribution are attributable to the singular component z_t, specifically the main harmonic component forcing the sea level to increase at certain tidal stages. The ACF, PACF, and periodogram analysis confirm this assumption (Figure 3.4). The analysis of the autocorrelation validates the persistence and cyclical nature of the process. The ACF shows a non-decaying sinusoidal pattern, confirming a singular process with

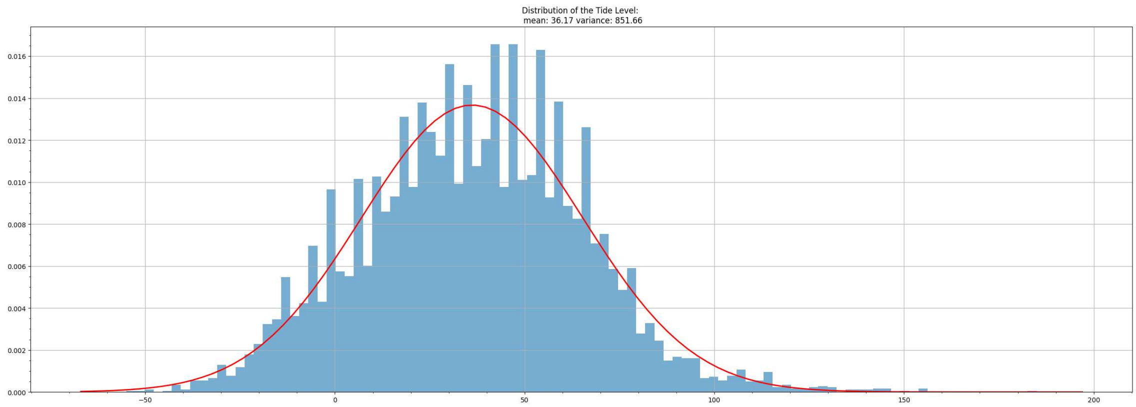

Figure 3.3: Empirical distribution of sea level observations (y_t) for the year 2019

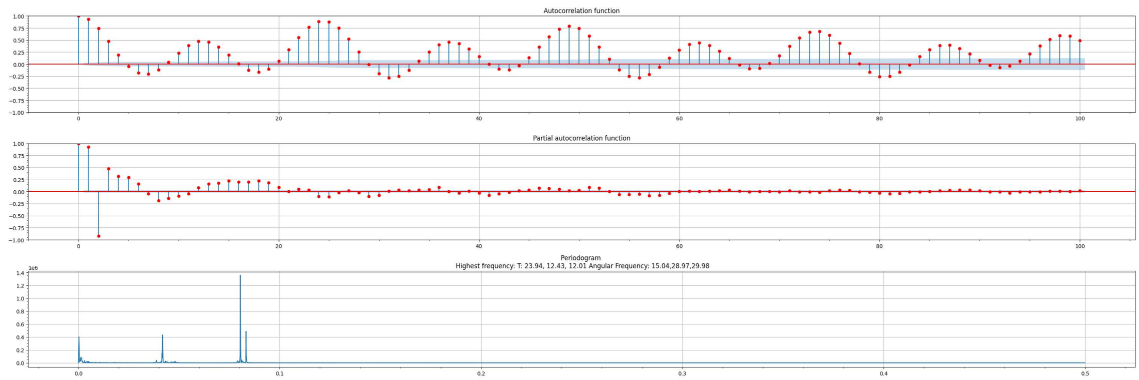

Figure 3.4: Autocorrelation (ACF), Partial Autocorrelation (PACF), and Periodogram analysis of the sea level y_t

infinite memory and suggesting a persistence factor ρ close to unity. Two main cyclical components emerge at lags 12 and 24. This is confirmed by the analysis of the periodogram, which shows very high sharp peaks, indicating a cyclical signal, for periods of 23.94 and 12.43, with angular frequencies of 15.04° and 28.97°. These frequencies correspond to the main diurnal (K1-lunisolar diurnal) and semidiurnal (M2-principal lunar semidiurnal) components of the astronomical tide as shown in Figure 3.1. This suggests a mixed cyclical pattern for the astronomical tide. This classification can be formalized using the Dietrich (1963) ratio:

$$F = \frac{A(K1)+A(O1)}{A(S2)+A(M2)} \quad (3.1)$$

Equation (3.1) classifies the tidal regime based on the ratio of diurnal to semidiurnal amplitudes: For $0 \leq F \leq 0.25$ the tide is semidiurnal, for $0.25 \leq F \leq 3$ the tide is mixed and for $F \geq 3$ the tide is diurnal. Using the values in Figure 3.1, the ratio for the Venice Lagoon is $F = 0.58$, confirming a mixed regime with dominant

semidiurnal components.
The periodogram also shows a high-power peak at very low frequency (0.01), indicating a long-term trend. This is consistent with the PACF, which shows a strong positive correlation at lag 1 followed by a negative one at lag 2, characteristic of a low-order autoregressive process. Together, these diagnostics confirm the mixed nature of the tidal signal.
As mentioned in section 3.1, transformations of the independent variables were used to represent the geological and physical relationship between these components and the water level.
Wind data consist of wind speed (m/s) and direction (N°). As illustrated in Figure 3.5, wind direction is characterized by a bimodal distribution. The peaks reflect the main wind on the Adriatic (Bora and Sirocco), which drive the surge events inside the Venice Lagoon. The wind speed is strictly positive with a heavy right skewness. To linearize these effects, wind is represented through the wind stress τ, defined by

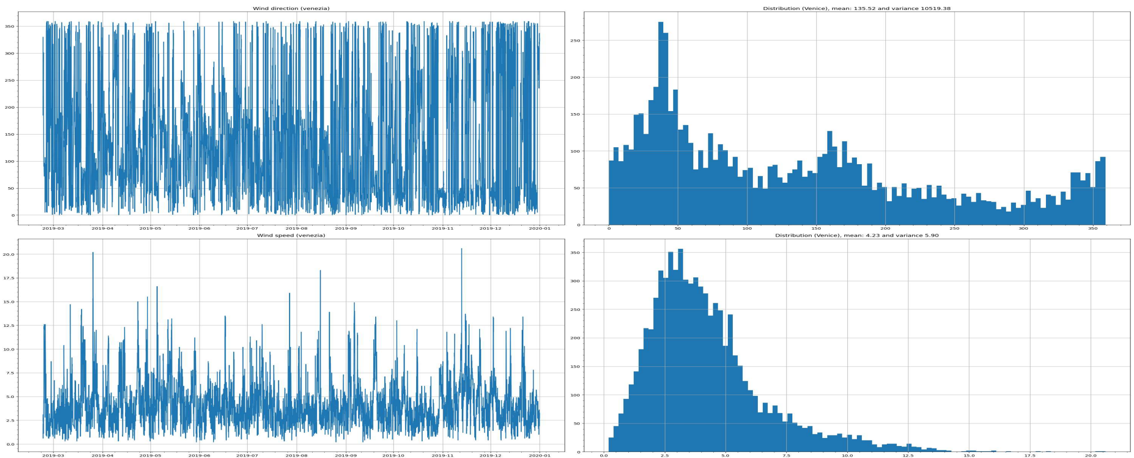

Figure 3.5: Empirical distribution of wind speed and direction observations in Diga Sud Lido for the year 2019

Large & Pond (1981) as a function of wind speed, potential temperature, and absolute humidity. Following the bulk aerodynamic method, the stress is approximated as a quadratic function of wind speed, decomposed into directional components as in equation (3.2):

$$\begin{gathered}\tau_{tx} = U^2 \cos(\theta) \\ \tau_{ty} = U^2 \sin(\theta) \\ \text{where } \theta \text{ is the wind direction in radians}\end{gathered} \tag{3.2}$$

As shown in Figure 3.6, the transformed components are centered around zero and take both positive and negative values. This allows the model to distinguish between wind movement pushing water toward (positive stress) the lagoon or pulling

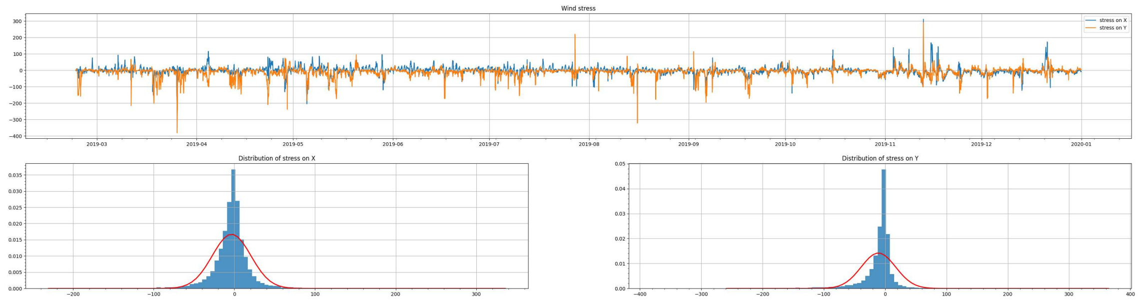

Figure 3.6: Empirical distribution of Wind Stress components

it away (negative stress). The transformation applied on circular data provides a linear representation of the effects. By squaring the value of the wind speed (U) we obtain a major impact for high values. Even if the majority of the observations are centered around 0, the histogram shows a high kurtosis, suggesting that the reference year contains extreme surge events.
Air pressure is decomposed into two physically distinct components. For the Venice Lagoon, the inverse barometric pressure (IBP) is computed as the deviation of each observation from the series mean. This variable captures the sea level response to pressure changes: lower pressure reduces the weight on the water surface, allowing sea levels to rise, while higher pressure suppresses them. As shown in Figure 3.7,

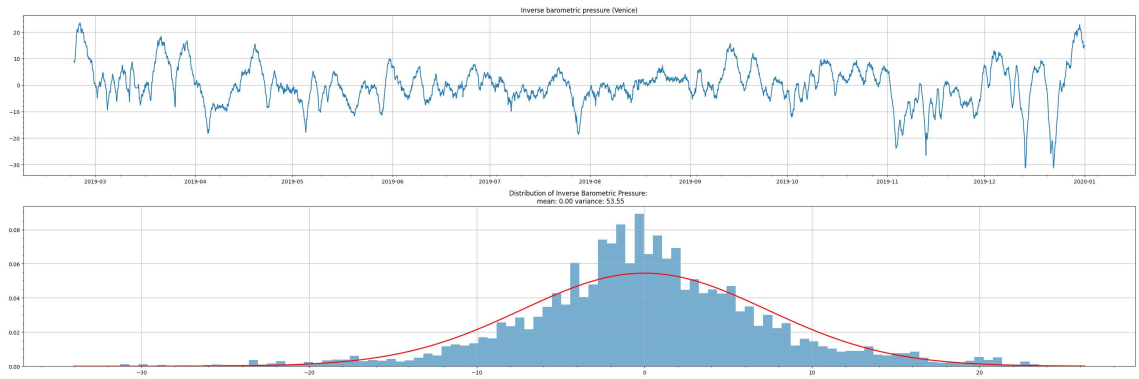

Figure 3.7: Empirical distribution of Inverse Barometric pressure at Punta della Salute

the IBP distribution is approximately Gaussian, with mean 0 and variance 53.55. Centering ensures that the model responds to pressure deviations rather than absolute values. For all other Adriatic stations (Figure 3.2), pressure is represented through the pressure gradients. This measure describes how the pressure differences between two geographical points affect the level of the tide. It is evaluated as the difference between constant pressure (measured in 1000 hPa) between Venice and various stations divided by their distance (equation 3.3).

$$PG = \frac{\Delta P}{\Delta D} \quad (3.3)$$

Distances are computed using the Haversine function. This metric measures the distance between two points given latitude and longitude considering the Earth's spherical curve.

$$a = \sin^2(\frac{\Delta\phi}{2}) + cos\phi_1 \cos\phi_2 sin^2(\frac{\Delta\lambda}{2})$$
$$c = 2\arctan(\sqrt{a}, \sqrt{1-a})$$
$$d = R * c$$

where R is the radius of the Earth.

The resulting pressure gradients are approximately normally distributed and exhibit small numerical magnitudes, typically averaging around 0.04 hPa/km (Figure 3.8). This is physically expected as the atmospheric pressure differential (Δp) is distributed over hundreds of kilometers along the Adriatic basin. The pressure

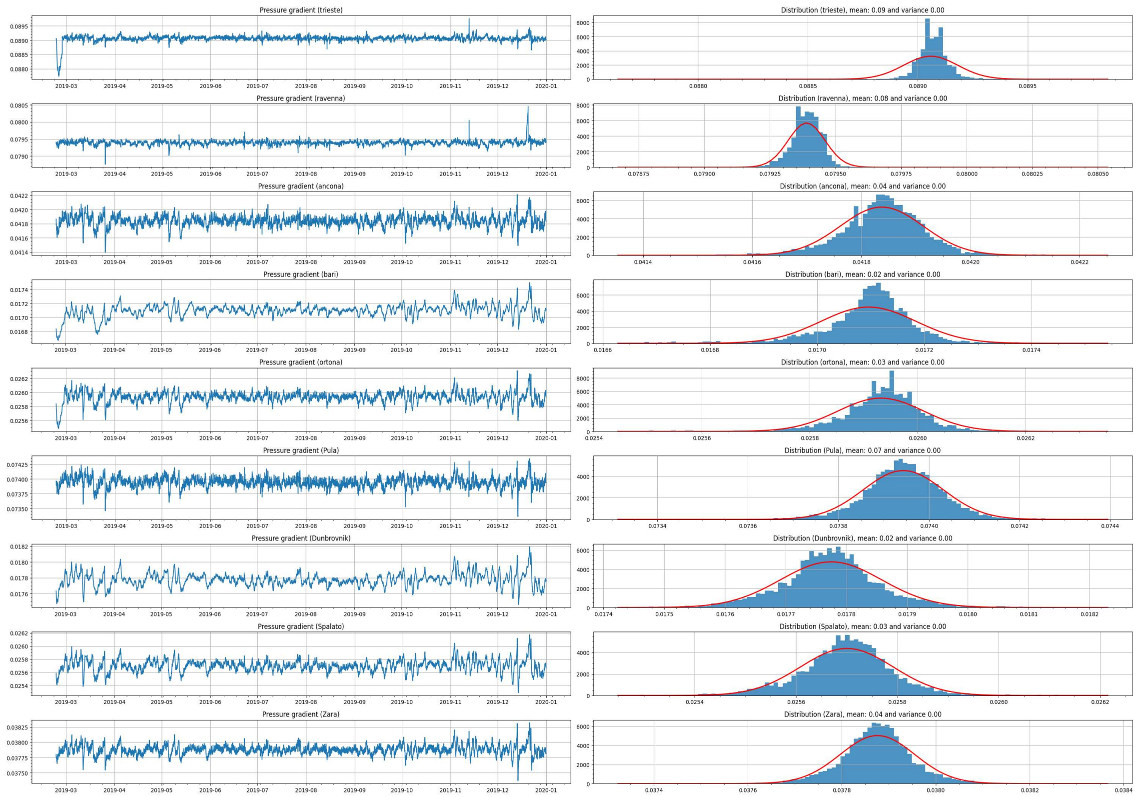

Figure 3.8: Empirical distribution of Pressure Gradient

transformations allow the model to capture both vertical sea level displacement, via the inverse barometric pressure, and the slope-driven effect of the pressure gradient along the Adriatic basin.

3.2. Model Specification

The empirical model was built following the Wold Decomposition Theorem applied by Sartore to the Venice tide (Sartore 1975). The approach consisted of a multi-step signal decomposition, where each component is estimated on the residuals of the previous one. After each element is subtracted from the observations, we can model the next one. For computation, the Kalman filter algorithm was employed, but each component enters the system in a different way.
The cyclical components were based on the eight primary harmonics ($M_2, S_2, N_2, K_2, K_1, O_1, P_1, S_1$). Each constituent defining the rotation of the cycle is expressed as follows:

$$\underbrace{\begin{bmatrix} \psi_t \\ \psi_t^* \end{bmatrix}}_{\alpha_t} = \rho \underbrace{\begin{bmatrix} \cos\lambda_c & \sin\lambda_c \\ -\sin\lambda_c & \cos\lambda_c \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} \psi_{t-1} \\ \psi_{t-1}^* \end{bmatrix}}_{\alpha_{t-1}} + \begin{bmatrix} \kappa_t \\ \kappa_t^* \end{bmatrix} \tag{3.4}$$

By considering the single harmonics as independent of one another, the general matrix for the SSM is designed as:

- Transition matrix T: 16×16 block-diagonal matrix. Each 2×2 block performs a trigonometric rotation for the eight primary harmonics (3.4).
- Measurement matrix Z: 1×16 vector $Z = [1, 0, 1, 0, \dots, 1, 0]$. This ensures that only the cosine term of each constituent is extracted.
- Disturbance Matrix R: This is a 16×16 identity matrix, ensuring that each stochastic shock κ_t affects each state component independently, thereby allowing for individual shifts in phase and amplitude
- Covariance Matrix Q: A 16×16 diagonal matrix where each diagonal pair corresponds to the variance of a specific tidal constituent. These variances are the hyperparameters estimated via Maximum Likelihood Estimation (MLE).
- Measurement Noise H: A scalar (1×1) representing $Var(\varepsilon_t)$. It captures the residual white noise not accounted for by the eight harmonics or the exogenous atmospheric variables.

The angular frequencies λ_c are treated as known constants derived from astronomical parameters (Figure 3.1). However, the physical expression of these cycles (amplitude and phase) can be influenced by shocks. To account for this, two distinct specifications were implemented:

- Deterministic Approach: The Covariance Matrix Q is set to zero. This assumes that the amplitude and phase of the tidal constituents are perfectly stable over time, effectively treating the singular component as a fixed harmonic regression.

- Stochastic Approach: The variances of the cyclical components (κ_t, κ_t^*) are estimated as non-zero hyperparameters. This allows the state vector α_t to evolve, granting the filter the flexibility to adapt the amplitude and phase to non-stationary shifts in the lagoon's dynamics.

To ensure the completeness of the singular components, a post-estimation diagnostic was performed on the residuals generated by the filter. Following the initial execution, a periodogram analysis was conducted to identify any remaining periodicities. Spectral peaks appearing at frequencies lower than 0.01 were attributed to long-term effects and trends. Consequently, only significant peaks with frequency values higher than 0.01 were identified as signals not captured by the initial primary harmonics.
The system was then iteratively refined by extending the SSM matrices: a 2×2 rotation block was added for each newly identified component, and the measurement and covariance matrices were expanded accordingly.
Once the structural complexity of the model was updated to include these latent oscillations, the filter was reiterated to achieve an optimal estimation where the remaining residuals align with a white noise process.
The initialization of the filter states is performed using distinct strategies for the primary harmonics and the newly identified unknown components. The eight principal harmonics are initialized with amplitudes and phases corresponding to the values provided by the Calendario delle Maree - 2019 (Figure 3.1), providing the system with a physically grounded deterministic baseline.
In contrast, the unknown parameters identified via periodogram analysis are initialized using a diffuse prior, with an initial state vector $\alpha = 0$ and an initial covariance matrix $P_0 = \kappa I$, where κ is a large scalar reflecting a diffuse prior. To account for the infinite memory and persistence of the cyclical components, a Local Linear Trend (LLT) was integrated into the system by means of first-order differencing. This ensures that the model captures non-stationary stochastic drifts in the lagoon's level, preventing long-term persistence from leaking into the residuals (3.5).

$$\begin{gathered} y_t = \begin{bmatrix}1\end{bmatrix} y_{t-1} + \begin{bmatrix}1\end{bmatrix} \eta_t \\ Z = [1],\, R = 1,\, Q = \sigma_y^2 \end{gathered} \tag{3.5}$$

To account for the long-term persistence and the periodic oscillations, the SSM was extended by stacking the matrices of the two independent processes. By treating these processes as independent within the augmented state vector, the model maintains the structural integrity of the Wold decomposition, while allowing the infinite memory of the trend to be isolated from the periodic singular components.
The parameters for the mean sea level (c) and the subsidence trend (α_1) were initialized using the values estimated by Sartore (1975), specifically 31 cm and 2.479 mm/year respectively. However, to account for potential variations in the 2019 dataset, these were treated as linear components within the system and estimated

by the filter via Maximum Likelihood Estimation (MLE). By allowing the filter to optimize these values, the model ensures that the baseline used in the innovation equation (3.6) is precisely calibrated to the specific sea-level height of the analyzed period. This calibration is performed before the singular cyclical components are isolated.

$$v_t = y_t - (Z\alpha_t + c + \alpha_1 t) \quad (3.6)$$

To ensure that varying scales across datasets do not bias the predictions, both the dependent and independent variables are normalized using Z-score transformation. Once the coefficients are evaluated, using multiple linear regression, they are rescaled to align with the original dimensions of the filter's input data. These rescaled coefficients are then incorporated as fixed parameters for the exogenous atmospheric variables within the innovation equation.
The comprehensive model is built considering the different single components into a single State Space framework. By stacking the independent transition matrix for all the cyclical components and the local linear level, the filter can account for deterministic astronomical forces, stochastic shifts, long-term subsidence, and atmospheric-pressure effects. The final innovation equation is expressed as:

$$v_t = y_t - (Z\alpha_t + c + \alpha_1 t + \textstyle\sum_{i=1}^{n} \beta_j X_{i,t}) \quad (3.7)$$

where $\beta_j X_{i,t}$ represent the rescaled coefficients of the multiple linear regression. The results, which are explored in the next chapter, show that residuals remaining from the model behave like white noise. This confirms that the model has captured all the possible signals from the data.

Chapter 4

Empirical results

The case study was conducted on a series of 8760 observations covering the year 2019. The analysis follows the two main approaches -deterministic and stochastic-outlined in the previous chapter.

4.1. Deterministic process

4.1.1. Model 1

The initial model was built by treating the eight principal tidal harmonics as deterministic processes, where the variance of each component was set to zero. To account for the long-term persistence of these tidal cycles, first-order differencing was applied to the data. The constant terms for the mean sea level and subsidence were initialized at 31 cm and 2.479 mm per year.
The results of this model, as shown in Figure 4.1, demonstrate that treating the harmonics as perfectly predictable processes is inadequate for signal decomposition. The training phase identifies cyclical components, yet a significant portion of the signal remains in the residuals. These residuals exhibit a persistent sinusoidal pattern, indicating that a purely deterministic approach cannot fully capture the signal. The ACF and periodogram confirm these findings: the ACF displays two distinct cycles every 24 lags, indicating a mixed pattern within the residuals that the model fails to capture.
Additionally, the periodogram reveals very high power at periods of 23.93 and 12.40 hours. This confirms that the selected deterministic components fail to capture the signal at these specific frequencies, where energy remains trapped in the error term instead of being correctly modeled. The results of Model 1 highlight the necessity of moving beyond a purely deterministic framework for the Venetian Lagoon. Because the tidal signal inside the Lagoon is heavily affected by meteorological surges and

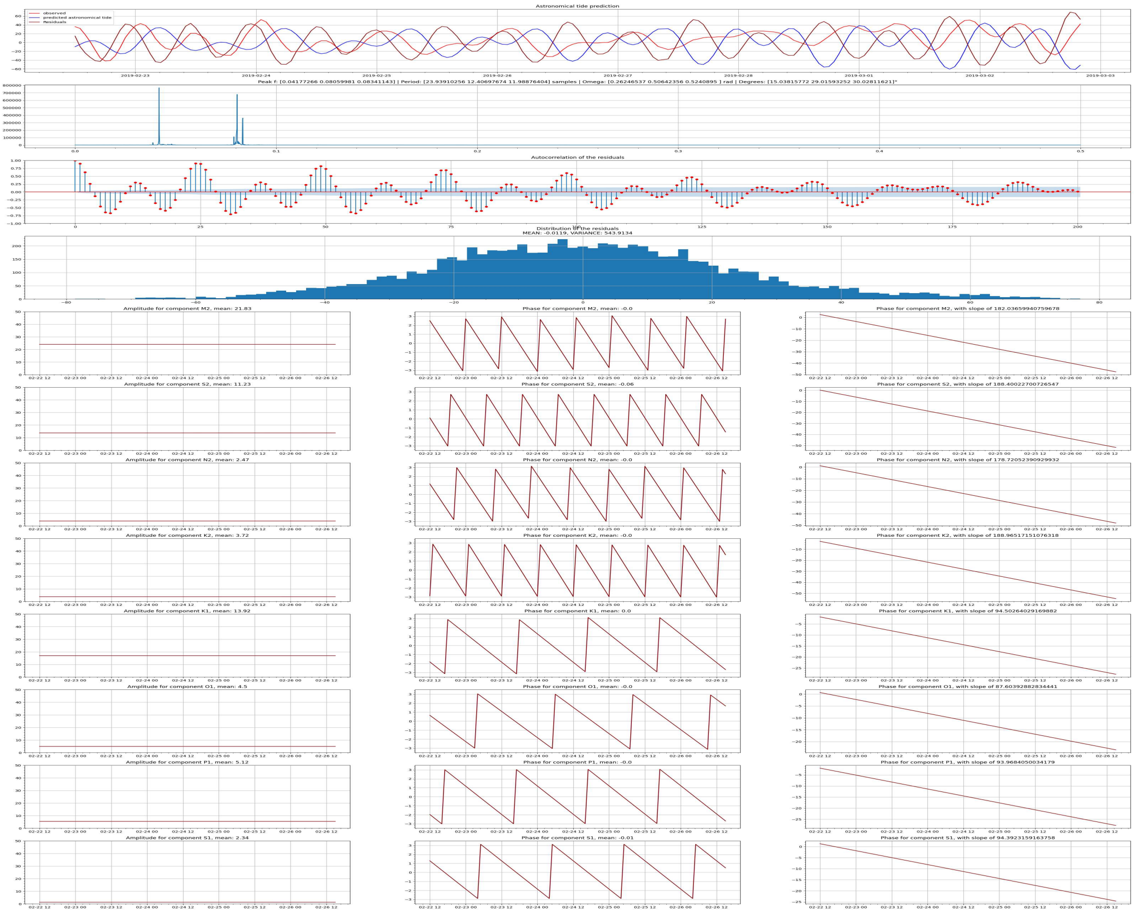

Figure 4.1: Model 1 - Visual analysis

varying water depths, the cycles cannot be treated as fixed parameters. By applying the State-Space Model (SSM) framework, the harmonics can instead be treated as stochastic hidden states.

4.1.2. Model 2

To refine the initial model, a second model is developed to absorb the remaining signal within the residuals. By analyzing the spectral peaks in the periodogram in Figure 4.1, it is possible to identify the specific frequencies that the initial model failed to capture. Consequently, two new cyclical components, "Unknown 1" and "Unknown 2", are integrated as hidden states into the SSM. These additional components are treated as stochastic cycles and initialized using a diffuse prior distribution. To represent the total uncertainty of the initial state of these components, the diagonal elements of the covariance matrix (P_0) corresponding to these states

are set to 10^6 and their initial value α_0 is set to 0.
The angular frequencies, which are treated as fixed terms, are derived from the spectral peaks identified in Figure 4.1. Specifically, they are set to 15.04° for the 23.93-hour period and 30.02° for the 12.40-hour period. Unlike the deterministic

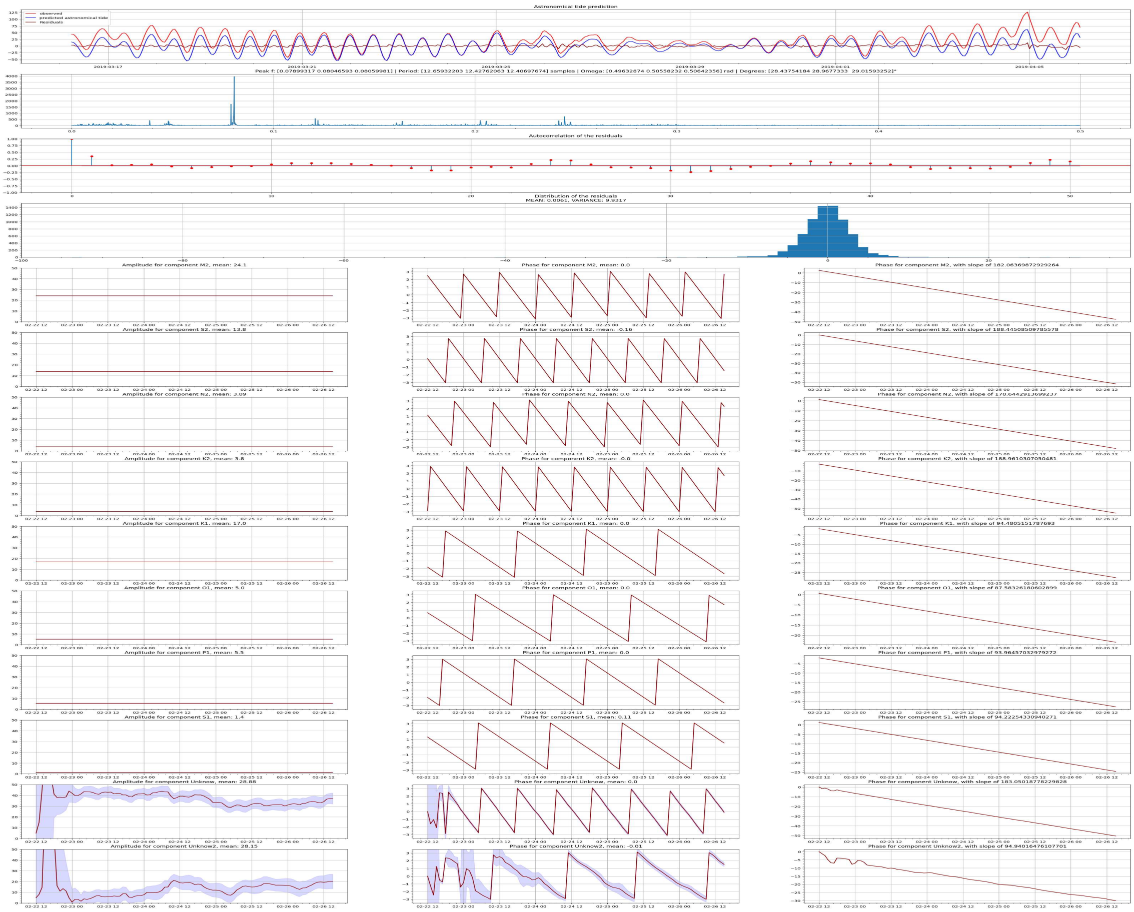

Figure 4.2: Model 2 - Visual analysis

components, these new states are treated as stochastic hidden states within the SSM. This allows the filter to estimate the amplitude and phase of these components as the system evolves, accounting for non-stationary shifts in the tide level and capturing the remaining cyclical effects of the residuals.
From the analysis of Figure 4.2, the cyclical effects remaining in the initial model were absorbed by the stochastic cycles. The histogram of the residuals is centered around zero, and the variance dropped significantly, from 543.91 in the initial model to 9.93. The training data plot shows that the predicted values accurately follow the cyclical pattern of the observed data, though a vertical offset remains. This offset represents the residual effect of unexplained atmospheric forces.
The ACF shows minimal residual cyclical patterns, confirmed by the periodogram,

where the highest peak (4,000) is significantly lower than the 800,000 observed in Figure 4.1.
The first stochastic cycle settles with an amplitude of approximately 35 cm, and its phase shows a semi-diurnal pattern, shifting from $-\pi$ to π every 12 hours. The second settles with an amplitude of around 20 cm, showing a diurnal pattern with a complete cycle every 24 hours. The constant slope of each phase after convergence suggests that the cycles follow a stable pattern over time, confirming that the filter has correctly identified the periodic structure of the tide.
This is further supported by the Kalman gain analysis in Figure 4.3. After the first 20 iterations, the model uncertainty drops to zero, reaching convergence. The deterministic components maintain a Kalman gain of zero throughout, as their parameters are fixed with zero variance, while the stochastic components and the estimated level stabilize once convergence is reached.

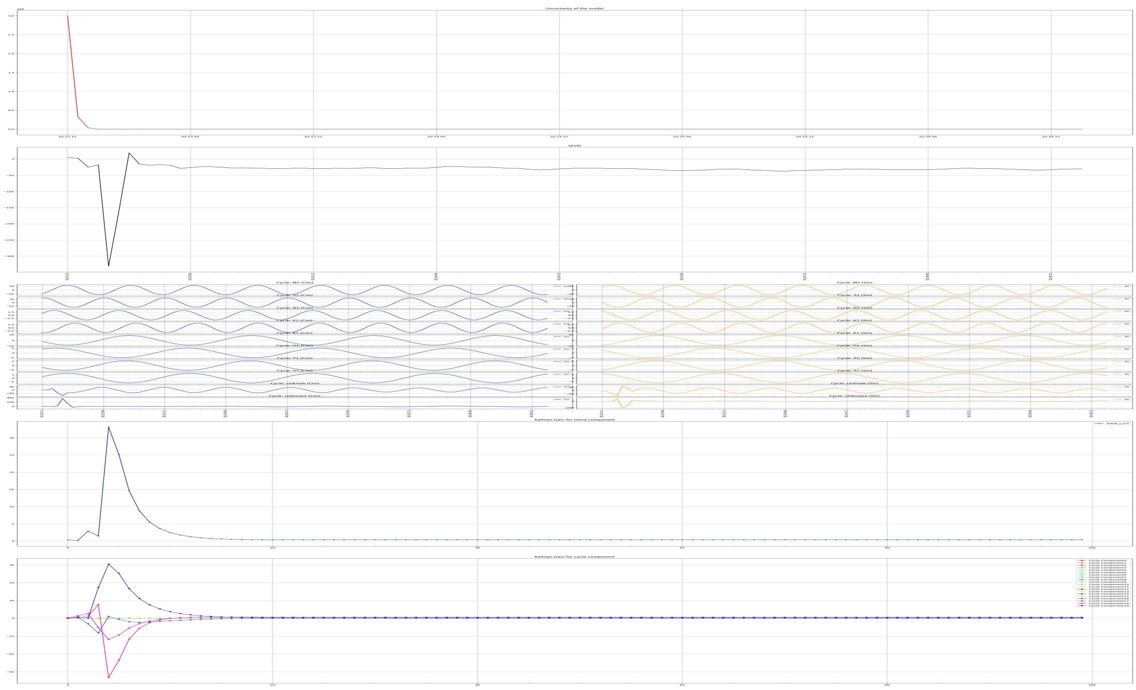

Figure 4.3: Model 2 - Kalman Gain analysis

The multi-step forecast results of Model 2, illustrated in Figure 4.4, visualize the model's forecasting performance across different horizons (24, 48, 72, and 168 hours). The blue lines represent the forecast and the red line represents the observations. The forecast follows the cyclical pattern of the tide but fails to absorb the variation introduced by exogenous effects. As the forecast horizon increases, the blue line diverges from the observations. As shown in Table 4.1, the evaluation metrics degrade significantly at greater horizons. The RMSE grows from 9.05 to 26.19 over one-week predictions, and the MASE value exceeds the threshold of 1 after the 24-hour fore-

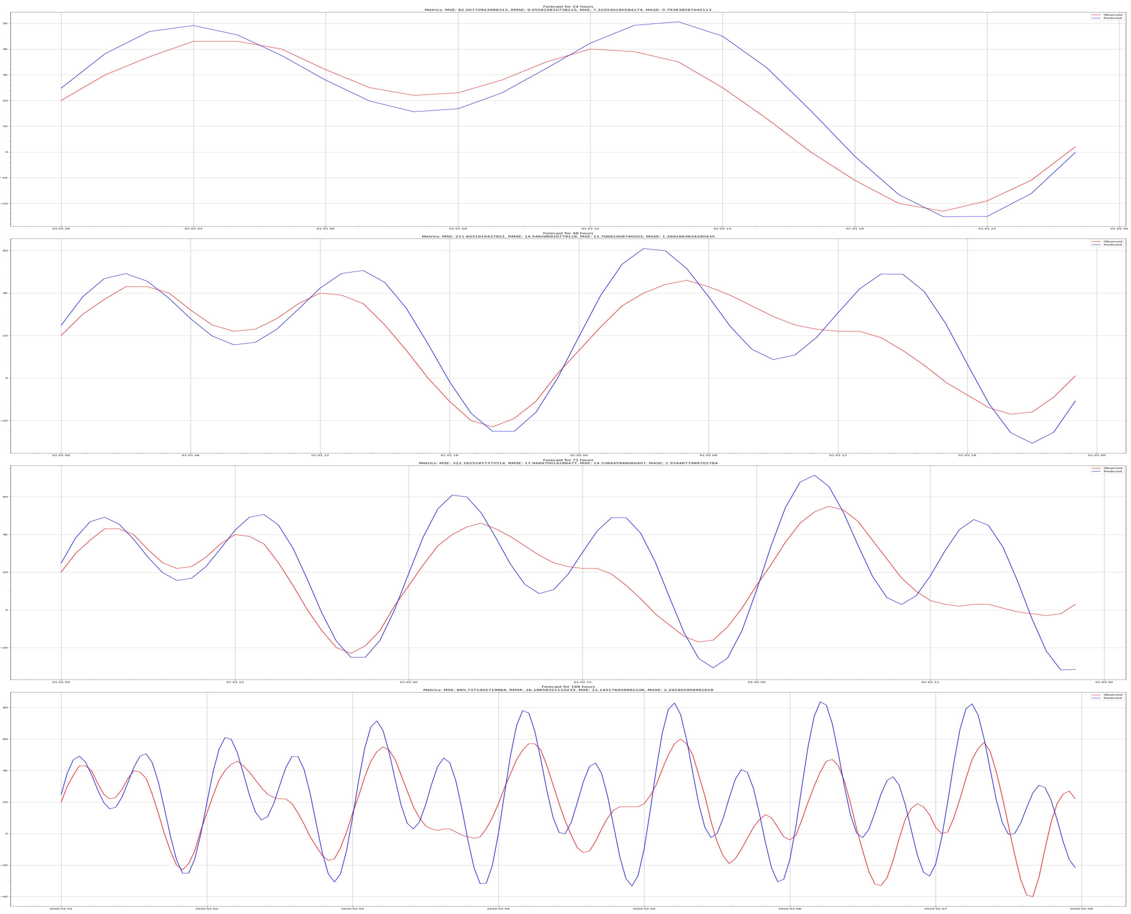

Figure 4.4: Model 2 - Forecast (24h, 48h, 72h, 168h)

cast, indicating that the model performs worse than a naive benchmark.
These effects can be explained by considering two main factors:

- Missing Innovation: During the forecast phase, the Kalman Filter cannot account for the process innovation because it lacks new observations. This increases the model's uncertainty, creating a visible divergence between the forecast and the observed data.

- Absence of Exogenous Effects: This model does not explicitly integrate exogenous effects, which consequently remain within the residuals of the training phase. As a result, the Kalman Filter attempts to adjust the hidden states of the cyclical components by treating these unexplained residuals as direct tidal effects. This causes the filter to misattribute the error of explainable components to the cycles themselves during the parameter update.

Metric	24 h	48 h	72 h	168 h
MSE	82.00	211.60	322.17	685.74
RMSE	9.05	14.55	17.95	26.19
MAE	7.32	11.70	14.34	21.14
MASE	0.79	1.27	1.55	2.30

Table 4.1: Model 2 Forecast Metrics

4.1.3. Model 3

As the residuals still contain unexplained signals, further decomposition is required. Following the methodology established for Model 2, the periodogram (Figure 4.2) reveals a cyclical persistence at the 12-hour period, with observed angular frequencies of 28.43° and 28.96°. To account for these components, two additional stochastic cycles have been integrated into the state model, initialized using a diffuse prior distribution, with constant angular frequency terms aligned with the corresponding periodogram peaks. The results, illustrated in Figure 4.5, demonstrates

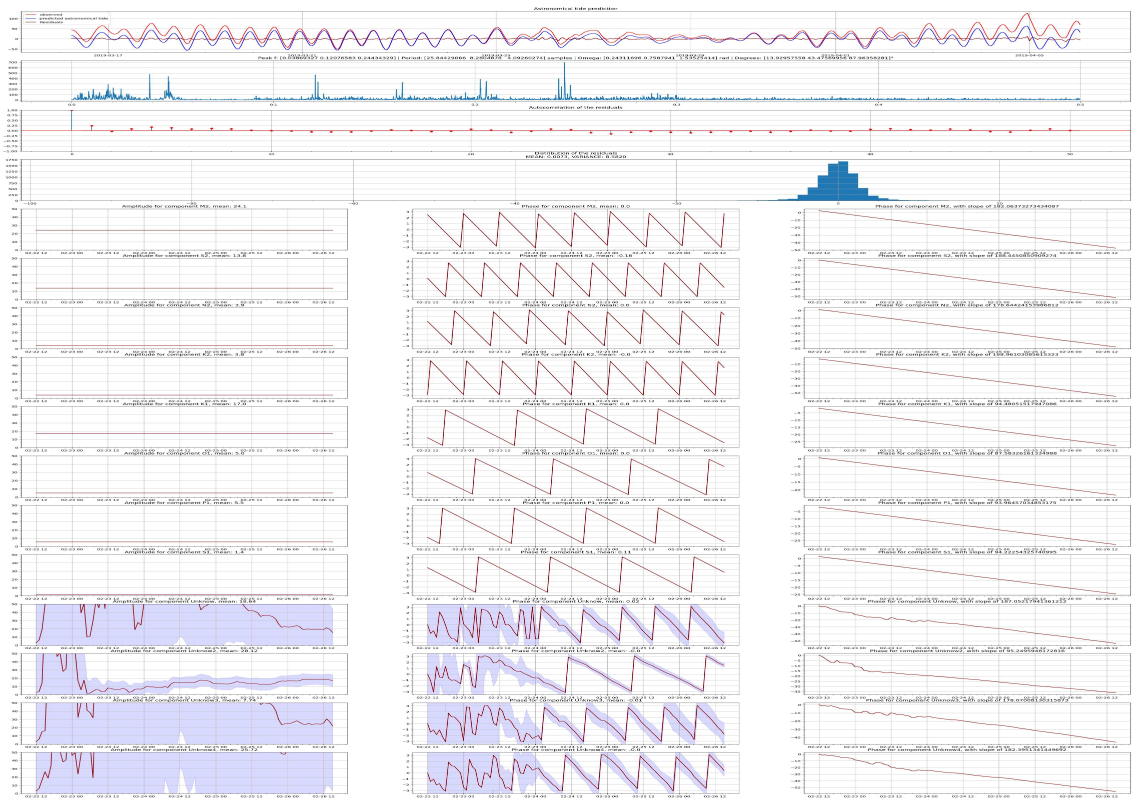

Figure 4.5: Model 3 - Visual analysis

that incorporating these two cyclical components absorbs the remaining signal, reducing the highest periodogram peak to 600. While some cyclical patterns remain, their residual magnitude is minimal.
The forecast results for Model 3, illustrated in Figure 4.6, cover horizons of 24, 48, 72, and 168 hours. The forecasted values match the observed pattern with much higher precision relative to previous results. As forecast horizons increase, a divergence remains, as expected: without new observations, the prediction error grows without the corrective adjustments provided by the innovation. The analysis of the

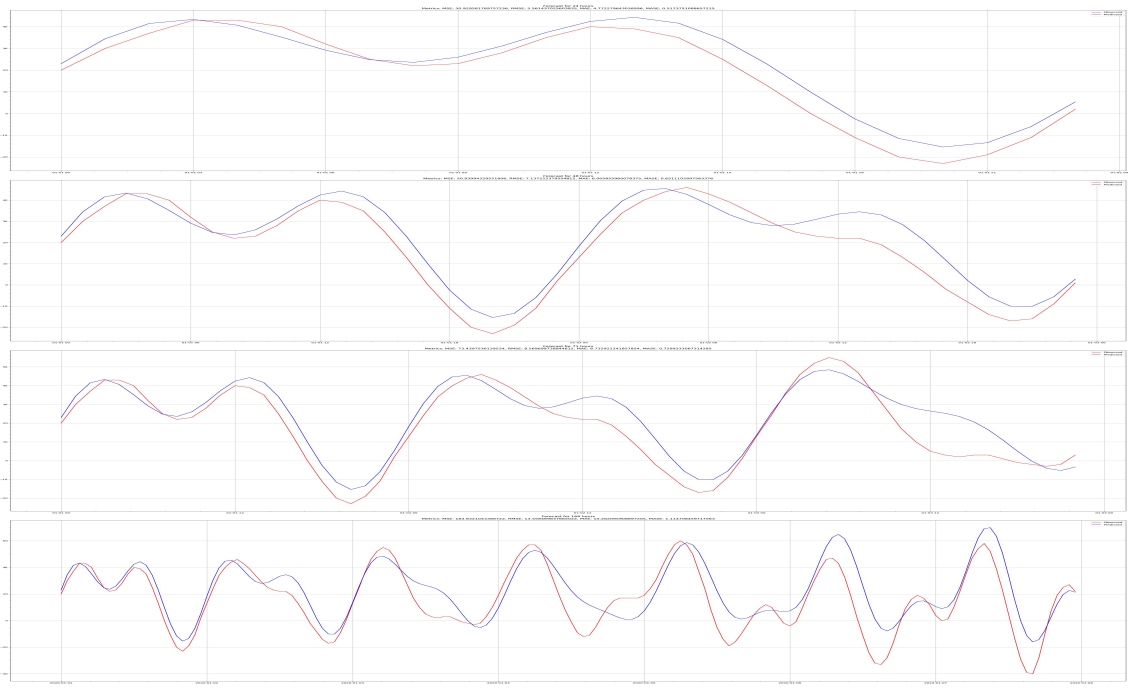

Figure 4.6: Model 3 - Forecast (24h, 48h, 72h, 168h)

Metric	24 h	48 h	72 h	168 h
MSE	30.93	50.94	73.43	183.83
RMSE	5.56	7.14	8.57	13.56
MAE	4.77	6.05	6.73	10.28
MASE	0.52	0.65	0.72	1.11

Table 4.2: Model 3 Forecast Metrics

forecast metrics in Table 4.2 demonstrates that the addition of the two stochastic components significantly improved the model performance:

- MSE (Mean Squared Error): Reduced across all forecast horizons. Specifically, the error reduction was 62% for 24 h, 75% for 48 h, 77% for 72 h, and 73% for 168 h.
- RMSE (Root Mean Square Error): Decreased for all forecast periods, with specific reductions of 38% for 24 h, 50% for 48 h, 52% for 72 h, and 48% for 168 h.
- MAE (Mean Absolute Error): Improved for all forecast periods, showing reductions of 34% for 24 h, 48% for 48 h, 53% for 72 h, and 51% for 168 h.
- MASE (Mean Absolute Scaled Error): Dropped below the 1.0 threshold for all forecasting periods, with the sole exception of the 168-hour forecast.

The results for Model 3 demonstrate that treating the eight principal harmonics as perfectly deterministic processes is not a suitable approach for signal decomposition, and the implementation of stochastic components to adapt the amplitude and phase of each component to the shocks inside the Lagoon is necessary.

4.1.4. Model 4

To quantify the impact of exogenous variables on the tidal series, a decomposition was performed on the observations after subtracting the evaluated cyclical components. The analysis was conducted on the residuals of Model 3, using a Multiple Linear Regression. The predictors used for the model are the following (the

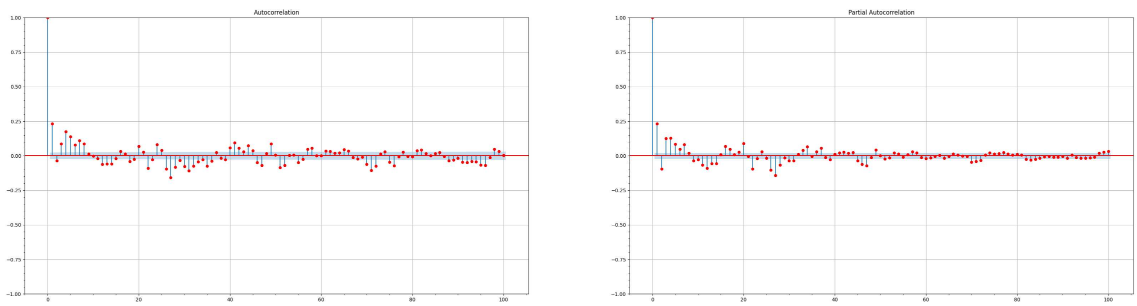

Figure 4.7: ACF & PACF of the residuals

evaluation of the single variables is explained in section 3.1.1):

- Inverse Barometric Pressure (IBP), measured in hPa, for Venice (Punta della Salute).
- Pressure Gradient, measured in hectopascals, for multiple stations on the Adriatic coast (Trieste, Ravenna, Ancona, Bari, Ortona, Pula, Dubrovnik, Spalato and Zara).

- Wind Stress component measured from data of wind speed and wind direction for Venice (Diga sud Lido).

To account for the non-instantaneous relationship between the predictors and the tide level, a temporal lag adjustment was applied to the dataset. The duration of this temporal shift was determined by the physical nature of each variable and its distance from the observation point. For instance, wind stress data were measured at Diga Sud Lido while the tide levels were sourced from Punta della Salute; a three-hour time shift was applied to compensate for the propagation delay between these two locations.
The time shifts for the pressure gradient observations were based on the propagation time of a gravity wave:

$$l = \frac{d}{\sqrt{gh}} \qquad (4.1)$$

where d represents the distance between Venice and the specified station, calculated using the Haversine formula, and $\sqrt{gh}$ represents the theoretical phase speed of a wave in shallow water. Differences in variable scales can introduce bias or instability

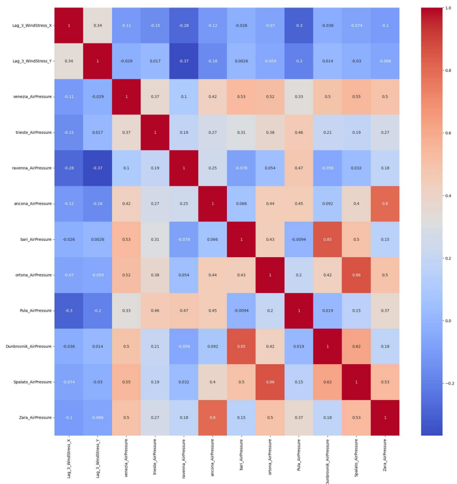

Figure 4.8: Correlation matrix of the independent variables

into the model. To ensure a uniform measurement basis, a Z-score standardization was applied to both the independent features and the dependent variable.
The correlation analysis (Figure 4.8) reveals a high degree of correlation, reaching values of 0.8 or higher, between specific variables. This is attributable to the geographical positioning of the stations: those situated within the same region of the Adriatic basin provide redundant information.
To address potential multicollinearity, highly correlated features were identified and excluded from the MLR, ensuring that the remaining variables provide meaningful, non-redundant information.

The MLR estimation results (Table 4.3 and Table 4.4) provide insight into the strength and limitations of assuming a linear relationship among the independent variables to explain the remaining signal. The most significant success of the model

Variables	coeff	coeff transf.	std err	P	95% CI
Wind Stress X	-0.0434	-0.0052	0.015	0.004	[-0.073, -0.014]
Wind Stress Y	-0.0271	-0.0028	0.015	0.074	[-0.057, 0.003]
IBP Venice	-0.1347	-0.0543	0.019	0.000	[-0.173, -0.097]
Pressure Gradient Trieste	-0.1053	-28.29	0.017	0.000	[-0.138, -0.073]
Pressure Gradient Ravenna	0.0787	32.61	0.016	0.000	[0.047, 0.110]
Pressure Gradient Ancona	-0.0592	-22.77	0.018	0.001	[-0.094, -0.025]
Pressure Gradient Bari	0.1310	42.97	0.018	0.000	[0.095, 0.167]
Pressure Gradient Ortona	0.0875	32.03	0.018	0.000	[0.052, 0.123]
Pressure Gradient Pula	-0.1405	-46.20	0.019	0.000	[-0.178, -0.103]

Table 4.3: MLR coefficients

R^2	$Adj - R^2$.	D.W	Jarque-Bera	Kurtosis	Cond. No.	AIC	BIC
0.067	0.065	2.027	11656453	234.56	3.14	1.469e+04	1.475e+04

Table 4.4: MLR regression diagnostics

can be assessed through its structural integrity and the white noise of the residuals:

- The Durbin-Watson statistic is equal to 2.027, which is close to its theoretical value of 2.0. This result indicates that the residuals do not exhibit any remaining autocorrelation between lags. This confirms that the residuals have reached white noise state.
- The Condition Number of 3.14 indicates no multicollinearity between the independent variables, confirming the effectiveness of the pre-processing phase. This ensures that each coefficient provides unique information, improving the numerical stability and reliability of the OLS estimation.
- Most estimated coefficients are statistically significant, with a p-value below 0.05

Although the structural stability of the model is solid, it is important to consider the limitations highlighted by the following metrics.

- The value of R^2 and adjusted R^2 are very low, around 6.5%. This indicates that the model can explain only a small part of the remaining variance. This

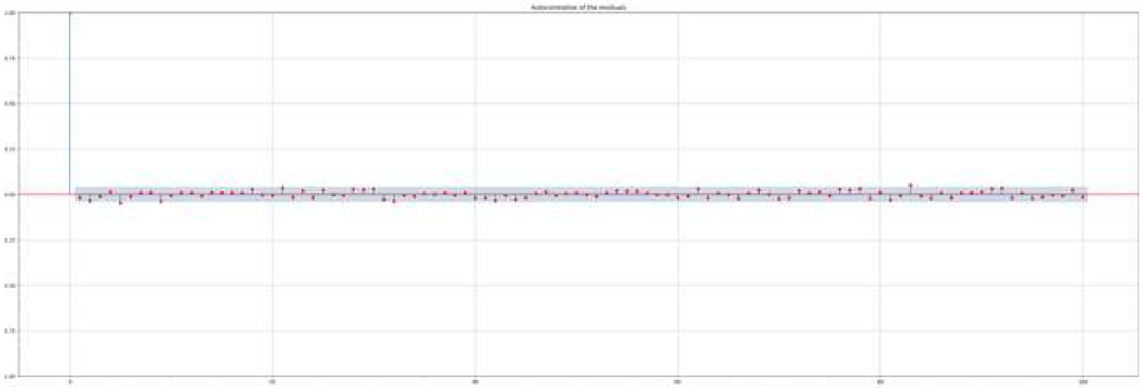

Figure 4.9: Autocorrelation function of the MLR residuals

suggests that after the primary harmonics have been extracted, the remaining signal is dominated by non-linear atmospheric noise

- The Jarque-Bera statistic has a very high value, indicating an extreme kurtosis. These "heavy tails" in the residual distribution are a direct consequence of the year 2019, which was characterized by historic storm surges and extreme shocks that exceeded the linear adaptive capacity of the model.

- The Jarque-Bera probability is 0, indicating that the residuals are not normally distributed.

- Both AIC and BIC show strong positive values (1.469×10^4), indicating that a strictly linear model may not be the most suitable approach for explaining the remaining complex effects in the data. These scores suggest that future refinements might require non-linear architectures.

The coefficients were rescaled after the evaluation:

$$\beta_{x,scal} = \beta \frac{\sigma_y}{\sigma_x}$$

The Multiple Linear Regression coefficients evaluated on the residuals of Model 3 provide a physical map of how the atmospheric effects affect the Venice Lagoon:

- Local Wind Stress: the values of the coefficients for the East-West (X) and North-South (Y) components are -0.0052 and -0.0028. These values are remarkably small, which means that by treating the cycles as stochastic—allowing their amplitude and phase to adjust based on shocks—the model absorbs the effects related to the wind component.

- Local Inverse Barometric Pressure: the coefficient for local air pressure in the Lagoon is -0.0543. While this value tends to be smaller than its theoretical effect of 1 cm/hPa, its sign respects the physical effect. For every increase in pressure, the tide level will drop by 0.0543 cm.

- Pressure Gradient: The coefficients for Bari (+42.97), Ortona (+32.03), and Ravenna (+32.61) are strongly positive, while Pula (-46.20), Trieste (-28.29), and Ancona (-22.77) are strongly negative. These coefficients measure how pressure in other locations within the Adriatic Basin pushes water towards the Lagoon. Although the magnitudes of the coefficients seem very high, they are suitable for the analyzed series given that the predictor variables themselves take very small numerical values.

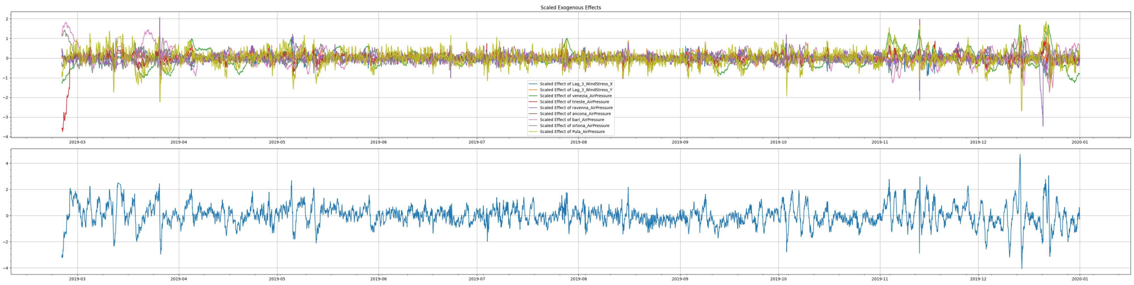

Figure 4.10: Tide Level variation of the independent variable

The coefficients are used as exogenous variables in the Kalman filter for the final model. They are applied to the series of each independent variable after subtracting their respective means. This transformation is performed to ensure that the model identifies changes in the water level based on the variation of the independent variables (Figure 4.10).
Considering the model as the combination of the previously applied multi-step decomposition, the structure is composed as follows:

- Eight deterministic cyclical components representing the main harmonics of the Venice Lagoon ($M_2, S_2, N_2, K_2, K_1, O_1, P_1, S_1$).
- Four stochastic cycles, representing the cyclical effects not captured by the main harmonics
- The estimated mean sea level.
- The long-term trend, representing the subsidence effect on Venice Lagoon.
- The effect of the Atmospheric components treated as exogenous effects with fixed coefficients evaluated using MLR.

The final results, shown in Table 4.5, reflect a slight reduction in forecast metrics compared to Model 3 in Table 4.2. This occurs because atmospheric adjustment uses fixed coefficients derived from MLR to explain the remaining residuals. While this

Metric	24 h	48 h	72 h	168 h
MSE	29.16	48.32	69.89	183.45
RMSE	5.40	6.95	8.36	13.54
MAE	4.61	5.79	6.56	10.25
MASE	0.49	0.62	0.71	1.11

Table 4.5: Model 4 Forecast Metrics

introduces a more rigid physical structure to the model's exogenous components, it provides a deterministic explanation for atmospheric drivers. In Model 3, these drivers were handled through internal state adjustments.
The final results demonstrate that the application of stochastic components via the Kalman filter allows the model to adapt the amplitude and phase of the waves to shocks of exogenous effects.
While the introduction of atmospheric components provides some improvement in the results, the use of fixed Multiple Linear Regression coefficients could be substituted in future iterations. Treating exogenous effects as hidden states within the SSM framework would allow the model to dynamically account for atmospheric variability.

4.2. Stochastic Cycles

A final model was analyzed using the available data. This model considers that the eight principal harmonics are not strictly deterministic processes; instead, they are treated as stochastic cycles where the amplitude and phase are allowed to change. The results shown in Figure 4.12 demonstrate that a stochastic representation of

Metric	24 h	48 h	72 h	168 h
MSE	49.13	91.68	135.54	237.81
RMSE	7.00	9.58	11.64	15.42
MAE	5.88	7.67	9.20	12.19
MASE	0.63	0.83	0.99	1.32

Table 4.6: Stochastic model - Forecast Metrics

the eight main harmonics allows the model to absorb nearly the entire cyclical effect in the data, without the need for additional cyclical components.
Comparing these results with Model 1 reveals a significant improvement. Treating the eight main harmonics as perfect deterministic processes fails to capture the

cyclical effects, leading to high correlation in the residuals and a variance of 543.91 (Figure 4.1). By contrast, the stochastic representation allows these harmonics to explain the majority of the signal, reducing the residual variance to 8.51.
This is further supported by the results of the periodogram and the autocorrelation function (ACF). The highest periodogram peaks show a power of 700, which is very small compared to the peaks in Model 1. Furthermore, the ACF shows only minimal remaining cyclical effects, confirming that the majority of the cyclical pattern can be explained using only the eight main harmonics. The forecast results (Figure 4.11 and Table 4.6) show that for forecast horizons of up to 72 hours, the model follows the cyclical pattern of the observation, as expected: without any adjustment from the innovation, the model struggles to adjust the tide level given increasing uncertainty.

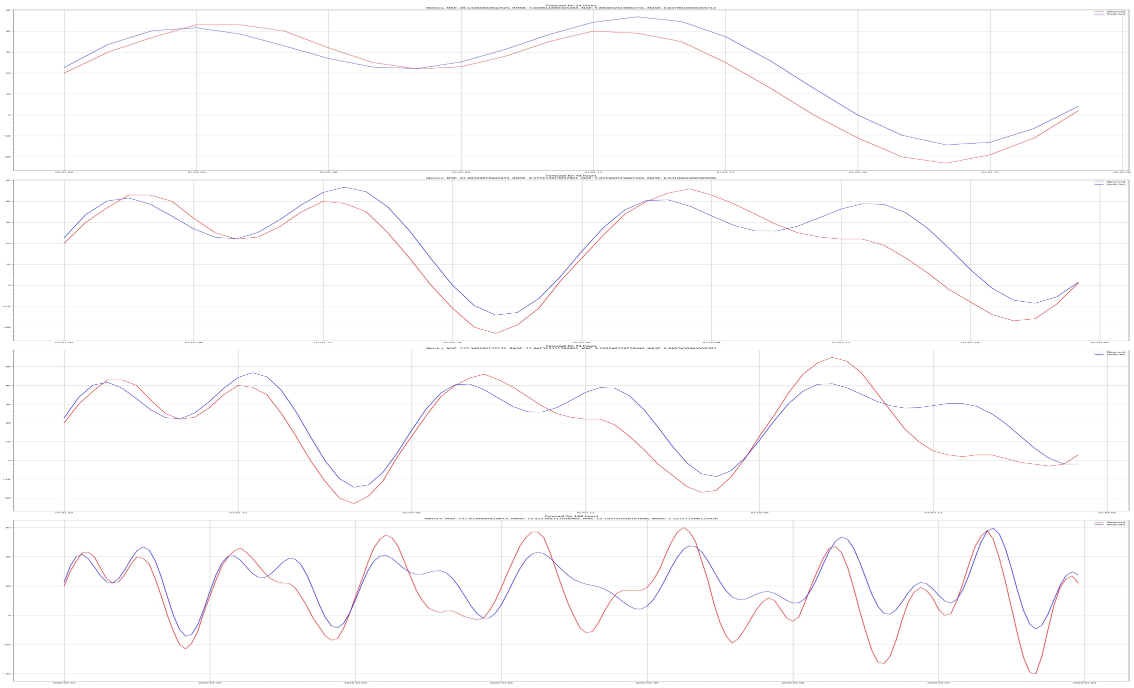

Figure 4.11: Stochastic Model - Forecast (24h, 48h, 72h, 168h)

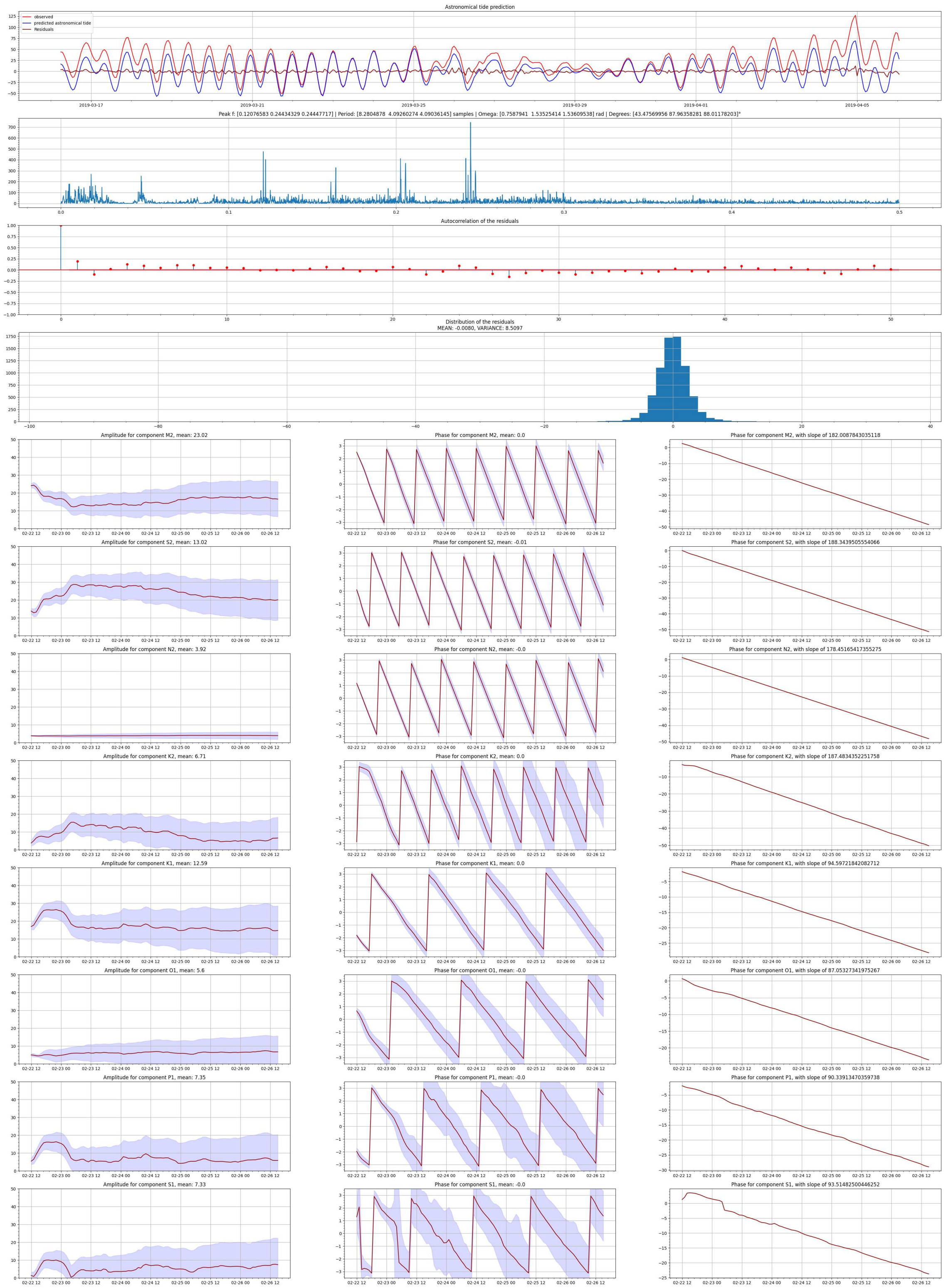

Figure 4.12: Visual Analysis for principal harmonics as stochastic components

Chapter 5

Conclusion

This thesis studies the feasibility of predicting sea levels in the Venice Lagoon using State Space Model (SSM) combined with the Kalman Filter. The study was motivated by the increasing frequency of Acqua alta events, which pose a threat to the cultural, artistic, and economic environment of Venice.
Building on the theoretical foundations of the Wold Decomposition Theorem and the study of Sartore (1975), the tidal signal was decomposed into its main components: an astronomical tide, a stochastic atmospheric surge, and a long-term geological drift driven by land subsidence and eustatism.
The empirical analysis, conducted on hourly observations from the year 2019 at the Punta della Salute station, provides several findings. The central result of this work is that treating the eight principal astronomical harmonics as perfectly deterministic processes is not a suitable approach for tidal signal decomposition in Venice Lagoon. As demonstrated by Model 1, a purely deterministic approach fails to capture the full cyclical structure of the tidal signal, leaving a significant amount of energy in the residuals.
This limitation arises because the tidal signal within the lagoon is subject to continuous perturbations from meteorological forcing, which cause the amplitude and phase of the astronomical harmonics to evolve over time.
The multi-step implementation of the model through the introduction of stochastic cyclical components, as implemented in Models 2 and 3, demonstrates improvements across all forecast horizons. By allowing the amplitude and phase of the tidal constituents to evolve dynamically through the Kalman Filter, the model was able to absorb the residual cyclical energy that a fixed harmonic regression cannot capture.
The final model (Model 4), which incorporates: deterministic cycles, stochastic cycles, and atmospheric exogenous variables estimated via Multiple Linear Regression, has reached a Root Mean Square Error of 5.40 cm for 24 hour forecast and 13.54 cm for 168 hour forecast.

This represents a significant improvement over the baseline deterministic process, outlined in Model 1. Furthermore, the pure stochastic model, which treated the principal eight harmonics directly as stochastic hidden states, confirmed that this specification is sufficient to absorb almost the entire cyclical signal. The residual variance dropped from 543.91 to 8.51 without the need for additional cyclical components.

The analysis of atmospheric components confirmed the hydrodynamic framework established by Vieira (1993). The pressure gradient coefficients provide a consistent picture on how atmospheric forces across the Adriatic Basin influence water level in the lagoon.

Stations in the southern Adriatic, such as Bari and Ortona, push water toward the lagoon (positive coefficients), while northern stations, such as Pula and Trieste, show negative coefficients and push water away from the lagoon. The inverse barometric pressure follows the expected theoretical sign, though its magnitude was smaller than the theoretical 1 cm/hPa response. This is consistent with the fact that the stochastic harmonic components absorb a part of the effect. The low R^2 of the Multiple Linear Regression (6.5%) suggests that, once the cyclical pattern is subtracted, the remaining signal is driven by non-linear atmospheric dynamics that cannot be explained by a linear model.

Different limitations of this study should be highlighted: First, empirical analysis was restricted to a single reference year, 2019, which was characterized by historically exceptional storm surge events. Although this provided a challenging test for the model, it also means that the results may not generalize to years with different meteorological conditions. Validation across multiple years with varying storm surge characteristics would strengthen the generalizability of the results.

Second, exogenous atmospheric variables were used in the model as fixed regression coefficients, which prevent the model from adapting dynamically to changes in atmospheric forcing. Finally, the application of first-order differencing to account for the long-term persistence of the cyclical components, while effective, introduces a modeling choice not suitable for the study of the lagoon dynamics.

Different directions for future research emerge from this work. The most natural extension would be to incorporate the atmospheric exogenous variables directly as hidden states within the State Space Framework, rather than as a fixed regressor. This would allow the Kalman Filter to update the effects of the atmospheric components dynamically as new observations become available.

In addition, the model could be extended to a multivariate framework that models sea levels at multiple stations within the lagoon. The integration of non-linear components, such as neural network based approaches, which have been explored in other studies on the Venice Lagoon, for the atmospheric residuals could also be considered.

In conclusion, this study demonstrates that the State Space Model framework, combined with the Kalman Filter, provides a flexible and feasible approach for tidal prediction in the Venice Lagoon. The key contribution of this work lies in the stochastic treatment of the principal astronomical harmonics, which allows the model to adapt dynamically to environmental shocks. These results confirm that this approach represents a solid step relative to the traditional deterministic harmonic regression.

Bibliography

[1] A, T. and P, C. (2011). Il modello stocastico per la previsione di marea a Venezia [The Stochastic Model for tide Forecasting in Venice]. Atti dell'Istituto Veneto di Scienze, Lettere ed Arti. Classe di scienze fisiche, matematiche e naturali, CLXIX (2010–2011). Istituzione Centro Previsioni maree, Comune di Venezia. Technical report.

[2] Bajo, M. (2010). La previsione di marea a Venezia: metodi statistici, deterministici e 4D-Var.

[3] Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1993). *Time series analysis forecasting and control.* 3 edition.

[4] Canestrelli, E., Canestrelli, P., Corazza, M., Filippone, M., Giove, S., and Masulli, F. (2007). Local Learning of Tide Level Time Series using a Fuzzy Approach. *IEEE International Conference on Neural Networks/IEEE ... International Conference on Neural Networks*, pages 1813–1818.

[5] Cartwright, D. E. (1988). Tides, Surges and Mean Sea Level–A handbook for engineers and scientists D. T. Pugh, John Wiley amp; Sons, Chichester, 1987, 472 pp, 55.50, 105.00.$GeophysicalJournalInternational$, 95(3) : 641.

[6] Di Nunno, F., Granata, F., Gargano, R., and De Marinis, G. (2021). Forecasting of extreme storm tide events using NARX Neural Network-Based models. *Atmosphere*, 12(4):512.

[7] Dietrich, G. (1963). *General Oceanography: An Introduction.* Interscience Publishers.

[8] Doodson, A. T. (1921). The harmonic development of the tide-generating potential. *Proceedings of the Royal Society of London Series A Containing Papers of a Mathematical and Physical Character*, 100(704):305–329.

[9] Durbin, J. and Koopman, S. J. (2001). *Time series analysis by state space Methods.* Oxford University Press.

[10] Geurts, M., Box, G. E. P., and Jenkins, G. M. (1977). Time Series Analysis: Forecasting and control. *Journal of Marketing Research*, 14(2):269.

[11] Hamilton, J. D. (2020). *Time Series analysis*. Princeton University Press.

[12] Harvey, A. (1993). *Time series models*. 2 edition.

[13] Large, W. G. and Pond, S. (1981). Open Ocean momentum flux measurements in moderate to strong winds. *Journal of Physical Oceanography*, 11(3):324–336.

[14] Orlic, M., Gacic, M., and Laviolette, P. (1992). The currents and circulation of the adriatic sea. *Institutional Archive of Ifremer (French Research Institute for Exploitation of the Sea).*

[15] Sartore, D. (1975). A statistical comparison of trends of the average monthly sea level in Venice and Porto Corsini (1947-1971) (in Italian).

[16] Umgiesser, G., Bajo, M., Ferrarin, C., Cucco, A., Lionello, P., Zanchettin, D., Papa, A., Tosoni, A., Ferla, M., Coraci, E., Morucci, S., Crosato, F., Bonometto, A., Valentini, A., Orlić, M., Haigh, I. D., Nielsen, J. W., Bertin, X., Fortunato, A. B., Gómez, B. P., Fanjul, E. A., Paradis, D., Jourdan, D., Pasquet, A., Mourre, B., Tintoré, J., and Nicholls, R. J. (2021). The prediction of floods in Venice: methods, models and uncertainty (review article). *Natural hazards and earth system sciences*, 21(8):2679–2704.

[17] Vieira, J., Føns, J., and Cecconi, G. (1993). Statistical and hydrodynamic models for the operational forecasting of floods in the Venice Lagoon. *Coastal Engineering*, 21(4):301–331.

[18] Wold, H. O. A. (1938). *A study in the analysis of Stationary Time Series*.

[19] Zanetti, R. and DeMars, K. J. (2013). Joseph Formulation of Unscented and Quadrature Filters with Application to Consider States. *Journal of Guidance Control and Dynamics*, 36(6):1860–1864.