Friedrich, Marina; Moussa, Karim; Shapovalova, Yuliya; van der Straten, David

Working Paper

Forecasting atmospheric ethane: Application to the Jungfraujoch Measurement Station

Tinbergen Institute Discussion Paper, No. TI 2025-025/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Friedrich, Marina; Moussa, Karim; Shapovalova, Yuliya; van der Straten, David (2025) : Forecasting atmospheric ethane: Application to the Jungfraujoch Measurement Station, Tinbergen Institute Discussion Paper, No. TI 2025-025/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/316191

TI 2025-025/III
Tinbergen Institute Discussion Paper

Forecasting Atmospheric Ethane: Application to the Jungfraujoch Measurement Station

Marina Friedrich[1]
Karim Moussa[2]
Yuliya Shapovalova[3]
David van der Straten[4]

1 Vrije Universiteit Amsterdam, Tinbergen Institute

2 Vrije Universiteit Amsterdam, Tinbergen Institute

3 Radboud University Nijmegen

4 Vrije Universiteit Amsterdam

Forecasting Atmospheric Ethane: Application to the Jungfraujoch Measurement Station*

Marina Friedrich[†1,2], Karim Moussa[1,2], Yuliya Shapovalova[3] and David van der Straten[1]

[1]Vrije Universiteit Amsterdam

[2]Tinbergen Institute

[3]Radboud University Nijmegen

Wednesday 9th April, 2025

Abstract

Understanding the developments of atmospheric ethane is essential for better identifying the anthropogenic sources of methane, a major greenhouse gas with high global warming potential. While previous studies have focused on analyzing past trends in ethane and modeling the inter-annual variability, this paper aims at forecasting the atmospheric ethane burden above the Jungfraujoch (Switzerland). Since measurements can only be taken under clear sky conditions, a substantial fraction of the data (around 76%) is missing. The presence of missing data together with a strong seasonal component complicates the analysis and limits the availability of appropriate forecasting methods. In this paper, we propose five distinct approaches which we compare to a simple benchmark – a deterministic trending seasonal model – which is one of the most commonly used models in the ethane literature. We find that a structural time series model performs best for one-day ahead forecasts, while damped exponential smoothing and Gaussian process regression provide the best results for longer horizons. Additionally, we observe that forecasts are mostly driven by the seasonal component. This emphasizes the importance of selecting methods capable of capturing the seasonal variation in ethane measurements.

*Author contribution statement: **Marina Friedrich**: Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing - original draft, review & editing. **Karim Moussa, Yuliya Shapovalova, David van der Straten**: Conceptualization, Investigation, Methodology, Software, Validation, Visualization, Writing - original draft, review & editing.

[†]Corresponding author: Department of Econometrics and Data Science, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV, Amsterdam, the Netherlands. E-mail address: m.friedrich@vu.nl.

1 Introduction

Atmospheric ethane is an indirect greenhouse gas, contributing to global warming. It belongs to the class of short-lived climate forcers which are broadly divided into methane and non-methane volatile organic compounds (NMVOC). They affect the climate and are often air pollutants (Szopa et al., 2021). Ethane is the most abundant NMVOC in the atmosphere, sharing important emission sources with methane – a major greenhouse gas with a high global warming potential (Franco et al., 2016). The main sources of ethane are anthropogenic (62% from leakage during production and transport of natural gas, 20% from biofuel combustion), while methane has both natural and anthropogenic sources (Xiao et al., 2008). This makes it hard to measure the fraction of methane released by the oil and gas sector. An estimate of this fraction can be provided with the help of ethane measurements (Visschedijk et al., 2018; Schaefer, 2019).

In addition to providing a better understanding of methane emissions from anthropogenic sources, the previous literature has identified two additional key reasons why it is relevant to study atmospheric ethane. First, ethane is an important precursor of tropospheric ozone. It contributes to the formation of ground-level ozone which is – unlike stratospheric ozone – a major pollutant affecting air quality. While ozone in higher levels of the atmosphere protects us from the sun's harmful ultraviolet rays, ground-level ozone damages ecosystems and has adverse effects on the human body (Fischer et al., 2014;

Franco et al., 2016). Second, since ethane and methane share the same oxidizer in the atmosphere, ethane influences the lifetime of methane. The oxidizer is the hydroxyl radical OH (Aikin et al., 1982; Rudolf, 1995). This implies that the more ethane there is in the atmosphere, the fewer OH radicals will be available for the degradation of methane, making ethane an indirect greenhouse gas (Collins et al., 2002). Its monitoring is therefore crucial for the characterization of air quality and the transport of tropospheric pollution. The main sources of ethane are located in the Northern Hemisphere, and the dominating emissions are associated with the production and transport of natural gas (Xiao et al., 2008).

Various time series of atmospheric ethane have been analyzed using econometric and statistical techniques. Friedrich et al. (2020a) analyze deterministic linear and nonlinear trends in the ethane burden above four measurement stations. Maddanu and Proietti (2023) study stochastic seasonality and trends in 15 time series of ethane. More trend analysis results can be found in Angelbratt et al. (2011), Franco et al. (2015) and Lutsch et al. (2020). Sun et al. (2021) analyze atmospheric ethane above Hefei in eastern China using Generalized Additive Models. In a recent study, Ortega et al. (2023) use exponential smoothing to predict business-as-usual values during the COVID-19 worldwide lockdown for various atmospheric gases, including ethane.

As this short review of the literature shows, the focus of previous studies has mostly been on understanding past developments. In this paper, our aim is to forecast the ethane

burden in the atmosphere. We focus on measurements obtained above the Jungfraujoch station in the Swiss Alps between February 1986 and July 2024, which is the longest time series of atmospheric ethane considered in the current literature. The Jungfraujoch location lies in the Northern Hemisphere where most emission sources are located. It is characterized by high dryness and low local pollution, leading to favorable measurement conditions (Franco et al., 2015). Nevertheless, around 76% of daily data is missing because measurements can only be taken under clear sky conditions. In addition, the data is characterized by a strong seasonal pattern since ethane degrades faster in summer than in winter. Together with the substantial amount of missing data, this poses a challenge for statistical analysis.

To address the above challenges, we select the following forecasting approaches which perform well in the presence of missing data and can take into account the seasonality of our data. The first approach is a fully deterministic trending seasonal model, which is one of the most commonly used models in the literature. It is used in Angelbratt et al. (2011), Franco et al. (2015), Lutsch et al. (2020) and Friedrich et al. (2020a) among others. We consider this model as a benchmark for comparison in forecasting. Second, in the same model framework, we use the method of discounted least squares which assigns higher weight to recent observations. Third, we employ damped exponential smoothing as presented in Gardner and McKenzie (1985). Fourth, we use a structural time series model in a state-space modeling framework as a classical and powerful method for time series

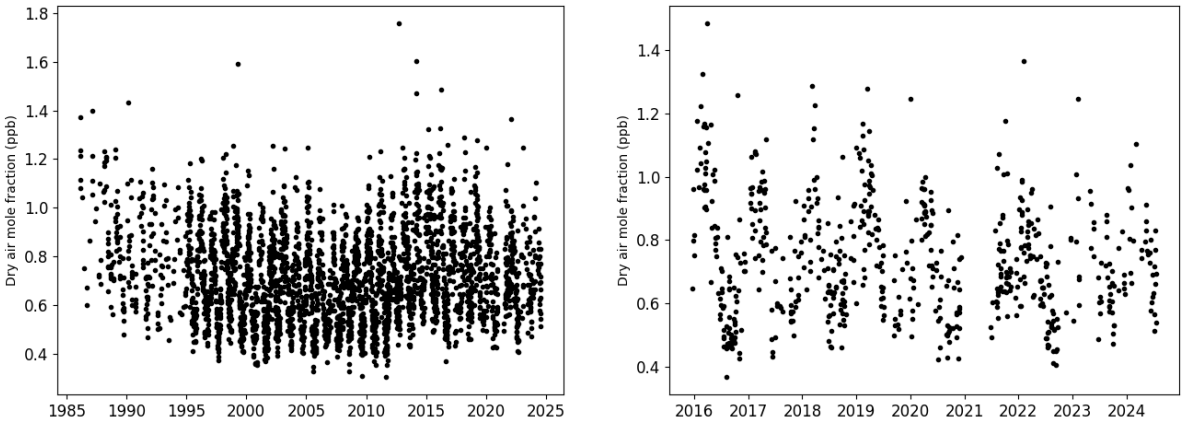
forecasting (Harvey, 1990). As a fifth approach, we use a local linear kernel regression which has previously been used in the literature for the analysis of past ethane trends (Friedrich et al., 2020a,b). The sixth approach is a Gaussian process regression, which can be seen as a Bayesian nonparametric regression technique. These forecasting methods have a natural way of dealing with missing data since no, or only minor adaptations are necessary. In addition, they can explicitly model seasonal variations.

Before we forecast the next three years of the atmospheric ethane burden above the Jungfraujoch measurement station, we compare the methods in an extensive forecasting evaluation exercise. We find that a structural time series model performs best for one-day-ahead forecasts, while damped exponential smoothing and Gaussian process regression provide the best results for longer horizons. Additionally, we observe that the forecasts are mostly driven by the seasonal component. This emphasizes the importance of selecting methods capable of capturing the seasonal variation in ethane measurements.

The remainder of the paper is structured as follows. Section 2 presents the ethane data and provides background. Section 3 discusses our forecasting methods. Section 4 presents the results of a pseudo out-of-sample forecast evaluation exercise and provides true out-of-sample forecasts. Section 5 concludes.

2 Data

We study time a series of atmospheric ethane obtained from a ground-based measurement station at Jungfraujoch located on the saddle between the Jungfrau and the Mönch, at 46.55°N, 7.98°W, 3580 meters altitude. Measurements are taken with the Fourier Transform InfraRed (FTIR) remote-sensing technique and contain ethane total columns as the number of molecules per cm^2 integrated between the ground and the top of the atmosphere. In this paper, we work with dry air mole fraction (DAMF) data in parts per billion (ppb) obtained from the Network for the Detection of Atmospheric Composition Change (NDACC). Compared to the non-normalized ethane total columns, the DAMF quantity includes information on surface pressure and water vapor total columns. The time series consists of 3257 observations of daily averages, ranging from February 1986 to July 2024. Although measurements are taken multiple times per day, they can only be taken under clear sky conditions, leading to a substantial number of missing observations: approximately 76% of daily observations are missing in our time series. Nevertheless, the measurement conditions at the Jungfraujoch station are favorable due to high dryness and low local pollution (Franco et al., 2015). This is the longest and currently most recent FTIR time series of ethane, with more than three decades of measurements available. Further details on the ground-based station at Jungfraujoch and on how measurements are obtained can be found in Franco et al. (2015).



(a) Full sample **(b)** Subsample

Figure 1: Ethane dry air mole fraction (DAMF) data in parts per billion (ppb). Panel (a) displays the full sample from February 1986 to July 2024, collected at the Jungfraujoch measurement station. The dataset covers 14,025 days, with 3,256 complete (non-missing) measurements, shown as dots. Panel (b) shows a subsample from December 2015 to July 2024. The subsample includes 3,135 days, with 652 complete measurements.

The data set is plotted in Figure 1. Panel (a) shows the full sample. To display features such as missing data more clearly, we additionally plot a subsample from December 2015 to the end of the sample in Panel (b). This will serve as our testing sample in the forecast evaluation exercise. In addition to missing data, the series displays a strong seasonal pattern, as ethane degrades faster in summer than in winter. Therefore, the measurements display local peaks every winter period. In summer, the atmospheric lifetime of ethane is at its minimum of around two months while it can be as high as ten months in the winter. On average ethane stays in the atmosphere around three months (Xiao et al., 2008; Helmig et al., 2016; Li et al., 2022). Previous analysis in Friedrich et al. (2020a) shows that the overall development of the series until 2019 has been characterized by two major

trend reversals. First, there has been a downward trend until around 2007, followed by an upward trend. Second, around 2015, the upward trend turned into a downward trend until the end of the series. The first trend reversal has been explained in the literature by increased activity in shale gas extraction in the United States (Vinciguerra et al., 2015; Franco et al., 2016; Helmig et al., 2016). The second reversal has been associated with a sharp decline in oil prices at the end of 2014. Lower oil prices affect the oil and gas industry and make shale gas extraction less profitable (Friedrich et al., 2020a).

3 Forecasting methods

Given the data characteristics described in the previous section and the modeling choices in the related literature, our forecasting methods are based on the following general model for the ethane measurements y_t

$$y_t = f(t) + \varepsilon_t \qquad t = 1, \ldots, n, \tag{3.1}$$

where the function $f(t)$ contains a seasonal and a trend component, and ε_t is the noise. In each forecasting method, the noise term is subject to different assumptions and we consider different specifications for the function $f(t)$. In most methods, it will be decomposed as

$$f(t) = \mu_t + s_t, \tag{3.2}$$

where μ_t models the long-term trend and s_t is the seasonal component.

3.1 A trending seasonal benchmark model

As a benchmark, we consider one of the most commonly used models in the related empirical literature on atmospheric ethane. It consists of a linear trend and a deterministic seasonal component. It has been used to analyze past ethane trends in, e.g., Angelbratt et al. (2011), Franco et al. (2015), Lutsch et al. (2020), and Friedrich et al. (2020a). A similar trending seasonal model has also been used to study temperature data in Diebold and Rudebusch (2022). The trend is specified as

$$\mu_t = \alpha + \beta t,$$

and the seasonal component follows

$$s_t = \sum_{j=1}^{S} a_j \cos(\lambda_j t) + b_j \sin(\lambda_j t), \qquad \lambda_j = \frac{2\pi j}{365.25}, \tag{3.3}$$

consisting of a combination of Fourier terms. The model is estimated by ordinary least squares. The number of Fourier terms is set to $S = 3$ in this paper whenever such a specification of the seasonal term is used. This is in accordance with the literature, which indicates that the seasonal variation in ethane measurements is well-captured by including three seasonal terms; see, e.g., Franco et al. (2015), Franco et al. (2016) and Friedrich et al. (2020a). In addition, Friedrich et al. (2020b) perform a frequency domain analysis to give more insight about the form of the periodic pattern present in the Jungfraujoch data. They find that including one term is clearly necessary and including up to three

terms further helps to model periodicity. Increasing S above three showed only minor effects. Three Fourier terms are also found to be sufficient for many other atmospheric ethane time series in Maddanu and Proietti (2023) who allow the seasonal component to vary over time.

3.2 Discounted least squares

By considering the trending seasonal benchmark model and assigning higher weights to recent observations, it becomes possible to forecast by extrapolating a local rather than a global trend. This concept lies at the heart of the discounted least squares (DLS) method for forecasting (Brown, 1963). Let $\theta = (\alpha, \beta, a_1, b_1, \ldots, a_S, b_S)'$ denote the parameter vector of the trending seasonal model, and consider the forecast function

$$f(t, h; \theta) = \alpha + \beta h + s_{t+h},$$

where h is the forecast horizon, and with the seasonal component s_t as defined in (3.3). In DLS, the parameters θ are estimated at each time t by minimizing the weighted sum of squares

$$Q_t(\theta; \omega) = \sum_{j \in \mathbb{I}_t} \omega^j \big\{ y_{t-j} - f(t, -j; \theta) \big\}^2, \qquad \mathbb{I}_t = \big\{ i \in \{0, \ldots, t-1\} \;\big|\; y_{t-i} \text{ is available} \big\},$$

for a given discount factor $\omega \in [0, 1]$, which controls the rate at which the weights of observations decrease with the distance in time. For $\omega = 1$ all observations are weighted equally, and we recover the trending seasonal from the previous section, while for $\omega < 1$ the

weights decay exponentially. The forecasting function is thus fitted by looking backwards from time t, and the objective function can be minimized analytically via weighted least squares. The resulting h-step ahead forecast at time t is

$$\widehat{y}_{t+h|t} = f(t, h; \widehat{\theta}_t), \qquad \text{with} \qquad \widehat{\theta}_t \in \arg\min_{\theta} Q_t(\theta; \omega).$$

The discount factor is a tuning parameter, and it is often recommended to use a value near one (e.g., Brown, 1963), since lower values of ω effectively reduce the sample size used in the regressions. This point is particularly relevant in our application, where the pervasiveness of missing data frequently leads to long periods without observations. To estimate the discount factor, we choose the value $\omega \in \{0.9999, 0.9995, 0.999, 0.99, 0.97, 0.94\}$ that minimizes the one-step ahead forecast mean squared error. We focus on the errors from the second half of the sample, as earlier forecasts, based on smaller samples, may not be representative of actual forecasting, where the full sample is utilized for prediction.

3.3 Damped exponential smoothing

Exponential smoothing offers another popular approach to forecasting, which is closely related to the method of DLS. We refer to Chapter 2.2 of Harvey (1990) for further discussion on their connection. Exponential smoothing, as originally presented in Holt (1957), has been applied to atmospheric time series in Ortega et al. (2023). The authors predict business-as-usual values for the COVID-19 period for various gases such as

atmospheric ozone, carbon monoxide, and ethane. Below, we consider a version of exponential smoothing with damping as proposed by Gardner and McKenzie (1985). To present the idea, we start by leaving aside the matter of seasonality and consider forecasting the deseasonalized data, y_t^d. The h-step ahead forecast according to the damped exponential smoothing (DES) method is

$$\widehat{y}_{t+h|t}^d = \alpha_t + \sum_{j=1}^{h} \phi^j \beta_t,$$

where $\phi \in [0, 1]$ is the damping parameter, and the level α_t and slope β_t are updated according to the following recursion:

$$\alpha_t = \lambda_\alpha y_t^d + (1 - \lambda_\alpha)\widehat{y}_{t|t-1}^d,$$

$$\beta_t = \lambda_\beta(\alpha_t - \alpha_{t-1}) + (1 - \lambda_\beta)\phi\beta_{t-1},$$

with smoothing constants $\lambda_\alpha, \lambda_\beta \in [0, 1]$. Lower values of λ_α correspond to a higher dependence of α_t on past observations, while lower values of λ_β yield more gradual updates of β_t. For $\phi = 1$ we obtain the original smoothing recursion of Holt (1957), and the forecast $\widehat{y}_{t+h|t}^d$ increases linearly with h; for values of $\phi < 1$, the additional contribution of the slope diminishes with the horizon, and the forecast converges to $\widehat{y}_{\infty|t}^d = \alpha_t + \beta_t \cdot \phi/(1 - \phi)$ as $h \to \infty$.

For the initialization, we proceed by analogy to Harvey (1990, p. 27) and account for the possibility of missing data by starting the recursion at time $t_2 + 1$, with t_2 being the

time index of the second available measurement. The previous level and slope are set to

$$\alpha_{t_2} = y_{t_2}, \qquad \beta_{t_2} = \frac{y_{t_2} - y_{t_1}}{t_2 - t_1},$$

with t_1 being the index of the first available measurement. To deal with missing data more generally, the recursion can be rewritten using the one-step ahead forecast error, $e_t^d = y_t^d - \widehat{y}_{t|t-1}^d$,

$$\alpha_t = \alpha_{t-1} + \beta_{t-1} + \lambda_\alpha e_t^d,$$

$$\beta_t = \phi\beta_{t-1} + \lambda_\beta\lambda_\alpha e_t^d.$$

This representation suggests setting $e_t^d = 0$ in the absence of y_t^d as a natural approach for dealing with missing data.

The usual way of handling seasonal effects in exponential smoothing is by introducing a seasonal variable s_t, which is updated according to $s_t = \lambda_s(y_t - \alpha_t) + (1 - \lambda_s)s_{t-k}$, where k denotes the period (Hyndman and Athanasopoulos, 2021, Ch.8.3). In our case $k = 365$, which would result in very infrequent updates of the seasonal effect for any given day of the year, an issue that is exacerbated by the large proportion of missing data, making this approach unsuitable for our application. We therefore proceed by estimating the seasonal effect separately. In particular, we use the trending seasonal model from Section 3.1 to estimate the seasonal term, $\widehat{s}_t$, then define the deseasonalized data as $y_t^d = y_t - \widehat{s}_t$, and apply the DES method to the time series $\{y_t^d\}$ as described above. The resulting

forecasts of the original data are obtained via

$$\widehat{y}_{t+h|t} = \widehat{y}^{d}_{t+h|t} + \widehat{s}_{t+h}.$$

To estimate the parameters λ_α, λ_β, and ϕ, we minimize the one-step ahead forecast mean squared error while imposing the damping parameter constraint $\phi \in [0.8, 0.98]$ (Hyndman and Athanasopoulos, 2021, Ch. 8), focusing on the predictive performance in the second half of the sample, as in the previous section.

3.4 A structural time series model

Structural time series models explicitly model the trend, seasonal, and noise components (Harvey, 1990). We will consider a structural time series model that can be expressed as a linear Gaussian state space model (SSM), which allows using the many established results in forecasting and parameter estimation based on the Kalman filter (KF; Kalman, 1960). This approach offers a natural solution for modeling and forecasting atmospheric ethane because the KF has an exact treatment of missing data (Harvey, 1990, Ch. 3).

For a time series of daily ethane measurements y_t, we specify the observation equation as in (3.1) and (3.2), where the errors follow $\varepsilon_t \sim \mathcal{N}(0, \sigma^2_\varepsilon)$ and are assumed independent and identically distributed. To ensure that the trend evolves gradually, it is modelled as an integrated random walk (Young et al., 1991), which is integrated of order two:

$$\mu_{t+1} = \mu_t + \delta_t, \qquad \delta_{t+1} = \delta_t + \eta^\delta_t, \qquad \eta^\delta_t \sim \mathcal{N}(0, \sigma^2_\delta).$$

The variable δ_t determines the direction of the trend; its specification as a random walk allows for the direction to change over time in a non-deterministic manner. To model the seasonal component, we use the following standard form (Durbin and Koopman, 2012, Ch. 3.2.2),

$$s_t = \sum_{j=1}^{S} a_{j,t} \cos(\lambda_j t) + b_{j,t} \sin(\lambda_j t), \qquad \lambda_j = \frac{2\pi j}{365.25}, \tag{3.4}$$

where λ_j implies $S \in \mathbb{N}$ different frequencies in a calendar year, and the coefficients follow autoregressive processes,

$$\begin{aligned} a_{j,t+1} &= \overline{a}_j + \phi_j(a_{j,t} - \overline{a}_j) + \eta_t^{a_j}, & \eta_t^{a_j} &\sim \mathcal{N}(0, \sigma_j^2), \\ b_{j,t+1} &= \overline{b}_j + \phi_j(b_{j,t} - \overline{b}_j) + \eta_t^{b_j}, & \eta_t^{b_j} &\sim \mathcal{N}(0, \sigma_j^2). \end{aligned}$$

This formulation enables the seasonal patterns to change over time, the importance of which has recently been highlighted by Maddanu and Proietti (2023). The parameters $\overline{a}_j$ and $\overline{b}_j$ control the unconditional means of $a_{j,t+1}$ and $b_{j,t+1}$, respectively. For the purpose of parsimony, the persistence parameter ϕ_j and scale parameter σ_j are shared by the j-th coefficients. The scale parameters are subject to the restriction $\sigma_\varepsilon, \sigma_\delta, \sigma_j \geq 0$, and for the persistence parameters $|\phi_j| < 1$ is imposed to separate the interpretation of the trend from the seasonal component. The model parameters are collected in the vector

$$\theta = (\sigma_\varepsilon,\ \sigma_\delta,\ \overline{a}_1,\ \overline{b_1},\ \phi_1,\ \sigma_1,\ \ldots,\ \overline{a}_S,\ \overline{b}_S,\ \phi_S,\ \sigma_S)', \tag{3.5}$$

which consists of $2 + 4S$ elements.

To formulate the structural time series model, consider the general linear Gaussian SSM with intercepts as in Harvey (1990),

$$\begin{aligned} y_t &= Z_t\alpha_t + d_t + \varepsilon_t, \qquad & \varepsilon_t &\sim \mathcal{N}(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + c_t + R_t\eta_t, \qquad & \eta_t &\sim \mathcal{N}(0, Q_t), \end{aligned} \tag{3.6}$$

for $t = 1, \ldots, n$, with y_t the measurement at time t, α_t the state vector, which is unobserved, ε_t and η_t Gaussian noise terms, and possibly time-varying system matrices Z_t, H_t, T_t, R_t, Q_t and vectors d_t and c_t of appropriate dimension. In our case, the state is

$$\alpha_t = (\mu_t,\ \delta_t,\ a_{1,t},\ b_{1,t},\ \ldots,\ a_{S,t},\ b_{S,t})',$$

with $N_\alpha = 2 + 2S$ elements, the state noise vector $\eta_t = (\eta_t^{\delta},\ \eta_t^{a_1},\ \eta_t^{b_1},\ \ldots,\ \eta_t^{a_S},\ \eta_t^{b_S})'$ consists of $N_\alpha - 1$ elements, and y_t and ε_t are scalars. For our structural model, only the $1 \times N_\alpha$ row matrix Z_t is time-varying,

$$Z_t = \begin{bmatrix} 1 & 0 & \cos(\lambda_1 t) & \sin(\lambda_1 t) & \ldots & \cos(\lambda_S t) & \sin(\lambda_S t) \end{bmatrix}.$$

The other system matrices, for which the subscript t will be omitted, are given by

$$d = 0, \qquad H = \sigma_\varepsilon^2, \qquad T = \begin{bmatrix} A & \mathrm{O} \\ \mathrm{O} & B \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix},$$

$$B = \mathrm{diag}(\phi_1,\ \phi_1,\ \ldots,\ \phi_S,\ \phi_S), \qquad Q = \mathrm{diag}(\sigma_\delta^2,\ \sigma_1^2,\ \sigma_1^2,\ \ldots,\ \sigma_S^2,\ \sigma_S^2),$$

$$c = (0,\ 0,\ (1-\phi_1)\bar{a}_1,\ (1-\phi_1)\bar{b}_1,\ \ldots,\ (1-\phi_S)\bar{a}_S,\ (1-\phi_S)\bar{b}_S)', \qquad R = \begin{bmatrix} 0'_{N_\alpha - 1} \\ I_{N_\alpha - 1} \end{bmatrix},$$

where O denotes the zero matrix of appropriate dimension, $\text{diag}(x)$ is a diagonal matrix having the vector x as its main diagonal, $I_{N_\alpha - 1}$ is the $(N_\alpha - 1) \times (N_\alpha - 1)$ identity matrix, and $0'_{N_\alpha - 1}$ is a $1 \times (N_\alpha - 1)$ row vector of zeros.

The linear Gaussian SSM is initialized by $\alpha_1 \sim \mathcal{N}(m_1, P_1)$ with initial mean vector m_1 and variance matrix P_1, the latter of which is usually assumed to be diagonal. For the state elements related to the seasonal component, a natural choice is to set the corresponding elements of m_1 to the unconditional means, $\overline{a}_j$ and $\overline{b}_j$, and the respective diagonal elements of P_1 to the unconditional variances, $p_j = \sigma_j^2/(1 - \phi_j^2)$. For the non-stationary elements μ_t and δ_t, it is standard to use diffuse initialization, which corresponds to letting $\kappa \to \infty$ in $\text{var}[\mu_1] = \text{var}[\delta_1] = \kappa$; in this case, the initial mean becomes irrelevant and is therefore set to zero (Durbin and Koopman, 2012, Ch. 5). In practice, this is often approximated by setting κ to a large number, say, $\kappa = 10^7$ (e.g., Harvey and Phillips, 1979). The resulting initial means and variances are

$$\mu_1 = (0,\ 0,\ \overline{a}_1,\ \overline{b}_1,\ \ldots,\ \overline{a}_S,\ \overline{b}_S)' \qquad \text{and} \qquad P_1 = \text{diag}(\kappa,\ \kappa,\ p_1,\ p_1,\ \ldots,\ p_S,\ p_S),$$

where it is assumed that the initial state α_1 is independent from the noise terms ε_t and η_t, and the latter are mutually and serially independent, as is standard in state space modeling (Durbin and Koopman, 2012, Ch. 3.1).

The SSM formulation enables forecasting the states via the KF, and the corresponding measurements can be predicted using the fact that $\text{E}[y_{t+h}|Y_t] = Z_{t+h}\,\text{E}[\alpha_{t+h}|Y_t]$. The

relevant formulae are discussed in Appendix A.1. Estimation of the model parameters θ is done by numerically maximizing the diffuse log likelihood described in Appendix A.2.

3.5 Local linear kernel regression

Former studies demonstrate that a non-parametric trend model estimated with local linear kernel smoothing is effective at modeling the ethane series (see, e.g., Friedrich et al., 2020a,b). We follow this approach and let the trend component of equation (3.2) be given by a smooth function of time $\mu_t = g(t/n) : [0, 1] \to \mathbb{R}$. While the seasonal component has been modelled by a combination of Fourier terms with constant coefficients in the previous literature, we extend this analysis by directly incorporating them into the model as regressors with time-varying coefficients. Hence, the seasonal component is given as in (3.4). This makes the combined model for y_t a time-varying coefficient model as considered in Robinson (1989) with $a_t = (a_{1,t}, \ldots, a_{S,t})'$ and $b_t = (b_{1,t}, \ldots, b_{S,t})'$. We adopt the common assumptions that $a_t := a(t/n)$ and $b_t := b(t/n)$ with $a(\cdot) = (a_1(\cdot), \ldots, a_S(\cdot))' : [0, 1] \to \mathbb{R}^S$ and $b(\cdot) = (b_1(\cdot), \ldots, b_S(\cdot))' : [0, 1] \to \mathbb{R}^S$ being smooth functions of time. We collect the regressors in $x_t := (1, \cos(\lambda_1 t), \ldots, \cos(\lambda_S t), \sin(\lambda_1 t), \ldots, \sin(\lambda_S t))'$ and the parameters in $\theta(\cdot) = (g(\cdot), a(\cdot)', b(\cdot)')'$. We obtain an estimate $\widehat{\theta}(t/n)$ for $t = 1, \ldots, n$ using an adapted version of the non-parametric local linear kernel estimator as in Cai (2007) and Friedrich and Lin (2024). The adaptation is necessary to account for missing observations. Details are given in Appendix A.3.

The fact that the trend and coefficient functions are expressed in terms of rescaled time (t/n) is a common assumption needed for consistency of the local linear estimator (Robinson, 1989). Estimation involves a kernel function and a bandwidth parameter ρ. Similar to Friedrich and Lin (2024), we employ the Epanechnikov kernel given by $K(x) = \frac{3}{4}(1-x^2)\mathbb{1}_{\{|x|\leq 1\}}$. Although bandwidth selection is an important aspect of our modeling approach, we first discuss how, for a given bandwidth ρ, the model can be used for forecasting. Next to this flexible trending seasonal model, we additionally consider an extended model which includes lags of the dependent variable. This has also been considered in the forecasting exercise of Friedrich and Lin (2024). In addition, Friedrich et al. (2020a) find that the ethane series exhibits strong autocorrelation, further motivating the use of lags. However, in our case, the inclusion of lags introduces an additional problem; missing observations in y_t now also affect the regressors. We thus add an additional component to the model, by including the most recent p observed values prior to time t as regressors in an extended model. To complete the construction of the regressors, we need to select the number of seasonal terms S and the number of lags p. As explained above, we set $S = 3$. Since the ethane data contains multiple periods with consecutive missing observations, including high-order lags can result in using observations far back in time. This is undesirable, as ethane measurements have a mean atmospheric lifetime of approximately 3 months. Thus, in the base model we set $p = 0$, and in the extended model we consider $p = 5$.

3.5.1 The non-parametric multistep ahead estimator

To forecast with this model, we employ a multistep ahead estimator proposed in Chen et al. (2018) for time-varying coefficient models. Specifically, the non-parametric multistep ahead estimate of y_{t+h} is given by the procedure presented in Algorithm 1. The non-parametric multistep ahead estimator involves iteratively performing one-step ahead forecasts which is shown in Chen et al. (2004) to perform better than a direct approach. After each forecast, we update the data by incorporating the forecast as if it were an observed value. Subsequently, we re-estimate the model and proceed to forecast the next value. By repeating this procedure h times, we obtain a sequence of forecasts $\widehat{y}_{t+1|t}, \ldots, \widehat{y}_{t+h|t}$. Note that for this procedure, we need the one-step ahead regressors x_{t+1} to be available. As mentioned in the previous section, the regressors x_t consist of deterministic seasonal terms and, in the extended model, lagged periods of y_t. Therefore, one can always obtain x_{t+1} when information up to time t is available.

Algorithm 1 Non-parametric multistep ahead estimator.

Step 1 Given the data up to time t, select the bandwidth ρ.

Step 2 Estimate the model using bandwidth ρ and obtain the estimated coefficients $\widehat{\theta}(\cdot)$, including $\widehat{\theta}(1)$, the local linear estimate at the right endpoint, using the observations up to time t.

Step 3 Obtain the one-step ahead forecast at time t,

$$\widehat{y}_{t+1|t} = x'_{t+1}\widehat{\theta}(1),$$

where x_{t+1} are the regressors at time point $t+1$.

Step 4 Update the in-sample observations $y_1, \ldots, y_t$ by including the pseudo observation $\widehat{y}_{t+1|t}$ as an estimate for y_{t+1}. Subsequently, update the corresponding regressors x_{t+1}. After updating, proceed as if we have observations up to time $t+1$, and let $t := t+1$.

Step 5 Repeat step 2 until 4 h times. The final one-step ahead forecast in Step 3 corresponds to $\widehat{y}_{t+h|t}$.

In Step 1 of the procedure, we must select the bandwidth ρ using the observed values up to time t. Selecting a suitable bandwidth is a challenging task that requires careful consideration. We employ a time series cross-validation approach, outlined in Appendix A.4. This procedure yields a bandwidth of $\widehat{\rho} = 0.41$ for the base model, and a bandwidth of $\widehat{\rho} = 0.48$ for the extended model. Since we update the data after each forecast, we could utilize the updated data to re-estimate the bandwidth for each one-step ahead forecast separately. Although this might increase forecasting accuracy, it imposes a great computational burden. Therefore, similar to Chen et al. (2018), the bandwidth selected in Step 1 is used for all the proceeding forecasts to reduce computational demand.

3.6 Gaussian processes regression

Gaussian process regression (GPR) is a non-parametric Bayesian regression technique where the function modeling the relationship between input and output values has a Gaussian process prior. It is given as in equation (3.1) with $f(t) \sim GP(0, k(t, t'))$ being a Gaussian process (GP) and $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ being Gaussian noise. Given a set of observed input-output pairs $D = \{(t_1, y_1), (t_2, y_2), \ldots, (t_n, y_n)\}$, the goal is to estimate the posterior distribution of the function $f(t)$. A GP is fully defined by its mean function, which in our case is assumed to be zero without the loss of generality, and covariance function specified by the kernel $k(t, t')$. The kernel function determines how the similarity between two inputs t and t' translates into their corresponding output values being correlated. The choice of the kernel affects properties such as the smoothness, periodicity, and generally the complexity of the functions the GP can model. Essentially, the kernel controls the GP's ability to generalize and capture patterns in the data.

3.6.1 Choosing the kernel function

In this subsection, we briefly describe the kernel functions used in our forecasting exercise for the ethane data. Various specifications of the kernel lead to different properties of the underlying function, such as smoothness, (non-)stationarity, and periodicity. Figure 3 illustrates three examples of GP samples with different kernel functions. Figure 3(a) illustrates samples from a GP with squared exponential function. This kernel depends on

two hyperparameters θ_1 and θ_2 through

$$k_1(t, t') = \theta_1^2 \exp\left(-\frac{(t-t')^2}{2\theta_2^2}\right), \tag{3.7}$$

where θ_1 controls the amplitude of the function (variation along y-asix) and θ_2 controls how wiggly the function is (larger values lead to slower changes and smaller values lead to faster local changes). In the context of time series this kernel is a natural choice for modeling long-term trends. Figure 3(b), illustrates GP samples with rational quadratic kernel given by

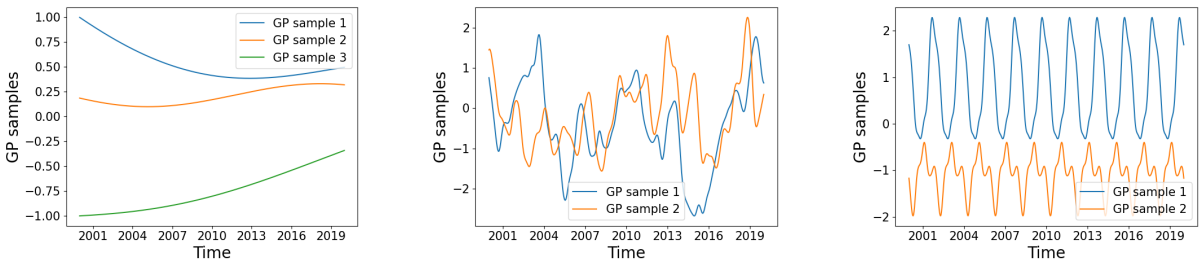
$$k_2(t, t') = \theta_3^2 \exp\left(1 + \frac{(t-t')^2}{2\theta_4\theta_5^2}\right)^{-\theta_4}, \tag{3.8}$$

with hyperparameters θ_3, θ_4, θ_5 and θ_6 (which describe magnitude, diffuseness and smoothness/variability). This kernel is often useful for describing medium-term trends and variations in the data.

Finally, periodicity can be accounted for with periodic kernel given by

$$k_3(t, t') = \theta_6^2 \exp\left(-\frac{(t-t')^2}{2\theta_7} - \frac{2\sin^2(\pi(t-t'))/\theta_9}{\theta_8^2}\right), \tag{3.9}$$

where θ_6^2 determines the magnitude of the functions, θ_7 allows for the decay-time of the periodic component and θ_8 determines the smoothness of the periodic component, θ_9 determines the period (and if it is known it can be fixed).



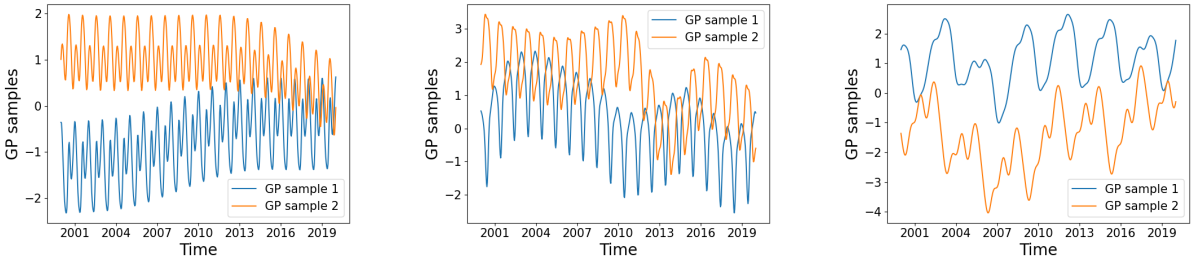
(a) **(b)** **(c)**

Figure 2: Samples from Gaussian process (GP) prior, different colours represent different GP samples. (a) Samples from a GP with squared exponential kernel; (b) Samples from a GP with rational quadratic kernel; (c) Samples from a GP with periodic kernel.

One of the properties that allows us to model more complex patterns in the data is that (certain) combinations of kernels lead to a proper kernel as well. In particular

$$k_f(t, t') = k_1(t, t') + k_2(t, t') + k_3(t, t'), \tag{3.10}$$

the sum of squared exponential, rational quadratic and periodic kernel would lead to behaviour which allows for all corresponding components: long-term trend, medium-range variations and periodicity. Figure 3(c) illustrates Gaussian processes samples from the combination kernel in Equation (3.10) with different hyperparameter values.



(a) **(b)** **(c)**

Figure 3: Samples from Gaussian process (GP) prior with combination of three kernels (squared exponential, rational quadratic and periodic), different colours represent different GP samples. In different sub-figures hyperparameters of the kernel are set to different values to demonstrate their effects. (a) The medium term component is irrelevant $\theta_5 = 10.$ in $k_2(t, t')$; (b) The medium term component is relevant θ_5 in $k_2(t, t')$; (c) Samples with longer period compared to (a) and (b).

3.6.2 Inferring hyperparameters of the kernel

Once we specified the kernel function, we fit the GPR to the data. The parameters of interest which we need to optimize (or infer using Bayesian inference) are θ_i, $i = 1, \ldots, 8$ when the kernel is specified according to the composition of 1) long-term, 2) medium-term and 3) periodic components as in equation (3.10). The hyperparameters are optimized using gradient-based optimization with the Gaussian log-likelihood as the objective function:

$$\log p(y|t, \theta) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2} \log |K_y| - \frac{N}{2} \log 2\pi, \tag{3.11}$$

where $K_y = K_f + \sigma_n^2 I$, is the kernel evaluated at training input points plus the variance of the observation noise.

3.6.3 Forecasting ethane time series with Gaussian process regression

In GPR, the predictive distribution at new input points is derived by conditioning the joint Gaussian distribution of observed data and predictions. The joint distribution of observed outputs y and the latent function values at the new input points f_* is given by

$$\begin{bmatrix} y \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K_f(t,t) + \sigma^2 I & K_f(t,t_*) \\ K_f(t_*,t) & K_f(t_*,t_*) \end{bmatrix} \right), \tag{3.12}$$

where $K_f(t,t)$ is the covariance matrix for the training data, $K_f(t,t_*)$ is the covariance matrix between the training data and test data, and $K_f(t_*,t_*)$ is the covariance matrix for test data and $\sigma^2 I$ is the noise variance.

Conditioning on the observed data (t,y), the predictive distribution for f_* is Gaussian

$$f_* \sim \mathcal{N}(\mu_*, \Sigma_*), \tag{3.13}$$

with

$$\mu_* = K_f(t_*,t)\big(K_f(t,t) + \sigma^2 I\big)^{-1} y, \tag{3.14}$$

$$\Sigma_* = K_f(t_*,t_*) - K_f(t_*,t)\big(K_f(t,t) + \sigma^2 I\big)^{-1} K_f(t,t_*). \tag{3.15}$$

When new observations are added, the predictive distribution can be updated efficiently without re-estimating the entire model by updating relevant terms in equations (3.14) and (3.15).

4 Results

This section first presents a forecasting experiment to evaluate the performance of the proposed methods across several horizons. Next, the best-performing methods are applied to perform true out-of-sample forecasting of atmospheric ethane in Section 4.2.

4.1 Forecasting experiment

To investigate the performance of the proposed forecasting methods, the following experiment is conducted for several fixed forecasting horizons h. We start by splitting the data into a training and a test sample, with the training sample comprising the first 80% of the complete (i.e., non-missing) measurements. More specifically, the training sample ranges from February 25$^{\text{th}}$, 1986 until December 19$^{\text{th}}$, 2015. The test sample, shown in Figure 1, starts the following day and ranges until July 19$^{\text{th}}$, 2024. Based on the original sample of length $n = 14025$ with 3256 complete measurements, the resulting training sample consists of 10890 measurements (2604 complete), and the test sample contains the remaining 3135 measurements (652 complete). Next, our forecasting methods are fit to the training sample. For a given horizon h, the fitted methods allow for computing the h-step ahead forecast at any time t conditional on the available data $Y_t = \{y_1, \ldots, y_t\}$. Denote this forecast by $\widehat{y}_{t+h|t}$. We compute these forecasts for all time points in the test

set and obtain the corresponding errors (if y_{t+h} is not missing)

$$e_{t+h|t} \equiv \widehat{y}_{t+h|t} - y_{t+h}$$

for $t = 1, \ldots, n-h$, such that the final forecast $\widehat{y}_{n|n-h}$ is a prediction of the last measurement, y_n. The forecast root mean squared error (FRMSE) is then computed by

$$\text{FRMSE}_h = \sqrt{\frac{1}{|\mathbb{I}_{\text{test}}|} \sum_{t \in \mathbb{I}_{\text{test}}} \left(e_{t|t-h}\right)^2},$$

with $\mathbb{I}_{\text{test}}$ denoting the set of time indices in the test set for which the measurements y_t are complete, and $|\mathbb{I}_{\text{test}}| = 652$ denoting the size of the test set.

Besides the trending seasonal model of Section 3.1, we consider two simple methods as naive benchmarks: the running sample mean and the random walk forecast. The latter sets $\widehat{y}_{t+h|t}$ equal to the last observed value at or before time t. For comparison, we use the DM test statistic of Diebold and Mariano (1995) with the heteroskedastic and autocorrelation-consistent estimator for the asymptotic variance of the test statistic.

Table 1: Forecast root mean squared error (FRMSE) of the forecast evaluation exercise.

	FRMSE									
h	1D	1W	2W	1M	2M	3M	6M	1Y	2Y	3Y
Simple benchmarks										
Mean	0.185	0.185	0.186	0.186	0.186	0.186	0.186	0.186	0.186	0.187
RW	0.167	0.183	0.185	0.199	0.219	0.262	0.307	0.193	0.188	0.188
TS	0.143	0.143	0.143	0.144	0.144	0.144	0.144	0.146	0.149	0.155
Proposed methods										
DLS	0.139	0.139	0.140	0.141	0.142	0.142	0.142	0.142	0.144	0.149
DES	0.139*	0.139*	0.139*	0.140*	0.140	0.140	0.140*	0.140*	0.140*	0.142*
KR	0.146	0.149	0.150	0.154	0.155^{-}	0.154	0.154	0.156	0.164^{-}	0.176^{--}
KRext	0.135*	0.144	0.147	0.150	0.152	0.151	0.151	0.152	0.158	0.165
SM	0.133**	0.141	0.144	0.150	0.158^{--}	0.164^{--}	0.168^{--}	0.177^{--}	0.204^{--}	0.232^{--}
GPR	0.137	0.137*	0.137*	0.138*	0.139*	0.141	0.140*	0.140*	0.140*	0.140**

FRMSE for various horizons h for the simple benchmarks (sample mean, random walk (RW), trending seasonal TS)) and the proposed methods: discounted least squares (DLS), damped exponential smoothing (DES), kernel regression (KR, the extended model (KRext) includes lags of the dependent variable), the structural model (SM) and Gaussian process regression (GPR). Entries with a (double) asterisk indicate that a method outperforms all benchmarks at a 10% (5%) significance level. Entries with a (double) minus sign show that the method is outperformed by one of the benchmarks at a 10% (5%) level. Shaded entries denote whether a model belongs to the 75% MCS.

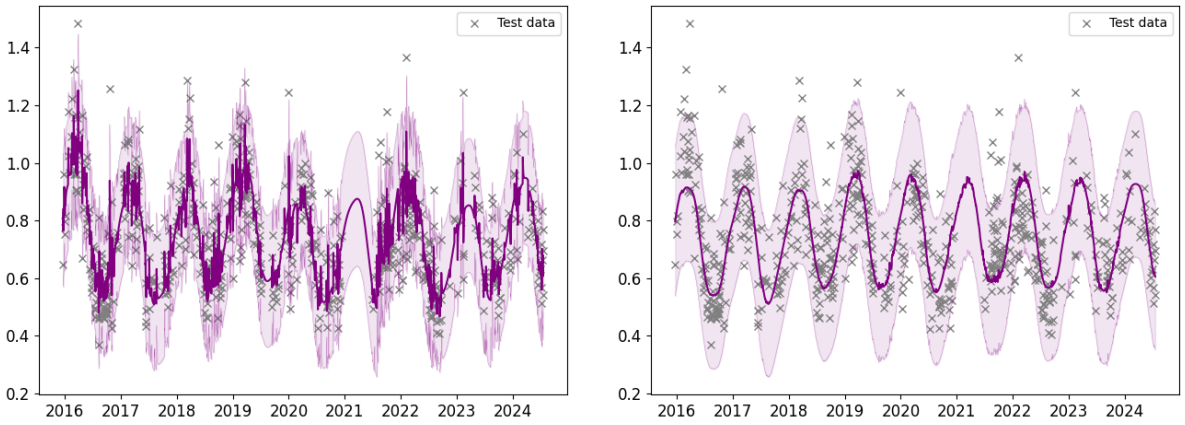
Table 1 presents the results of the forecasting experiment. The FRMSE is reported for horizons of one day (1D), one and two weeks (1W, 2W), one, two, three, and six months (1M, 2M, 3M, 6M) – where a month is taken as equivalent to 30 days – and one, two, and three years (1Y, 2Y, 3Y), with a year corresponding to 365 days. The first three rows present our benchmarks while the rows below the horizontal line show the results for our proposed methods. In addition, we indicate results of the DM tests in the following way. Entries with a (double) asterisk indicate that a method outperforms all benchmarks at a 10% (5%) significance level. Entries with a (double) minus sign mean that the method is outperformed by one of the benchmarks at a 10% (5%) level. Shaded entries denote

whether a model belongs to the Model Confidence Set (MCS) of Hansen et al. (2011) for a given horizon at a significance level of 25%, resulting in 75% Model Confidence Sets. For this procedure, we use the implementation of Bernardi and Catania (2018) and consider a squared loss function.

We observe that the simple trending seasonal model performs best out of all benchmarks. The DLS outperforms the trending seasonal for all horizons, but the difference in predictive accuracy is not significant. The performance of DES is similar to DLS in terms of the FRMSE at short horizons. At longer horizons, it does slightly better. The DM tests indicate that it outperforms the benchmarks at the 10% level for most horizons. The kernel regression performs better in the extended version, which includes lags of the dependent variable. It significantly outperforms the benchmarks at the 10% level for one-day ahead forecasts. The kernel regression without lags (base model) gets outperformed by the trending seasonal model at 2M, 2Y and 3Y horizons. The structural model performs best in terms of one-day ahead forecasts according to the FRMSE, and it outperforms the trending seasonal benchmark at the 5% level. For longer horizons the performance degrades, and starting at 2M it is outperformed by the trending seasonal model. GPR results are more stable in terms of the FRMSE. It performs particularly well for longer horizons, where it significantly outperforms the trending seasonal benchmark at the 5% level.

The 75% Model Confidence Sets closely align with the results above. For horizons

up to and including 1M, the MCS merely excludes the naive benchmarks and the kernel regression (base model). For the 2M, 3M, and 6M horizons, the MCS additionally excludes the structural model and the extended kernel regression. Following that, the 1Y horizon MCS excludes the trending seasonal model and contains the methods of DLS, DES, and GPR, while the 2Y horizon MCS consists of DES and GPR. Lastly, the 3Y horizon MCS only includes the GPR model. These results indicate that, as the horizon increases, the Model Confidence Sets become progressively smaller, only including models that are relatively stable in terms of the FRMSE.



(a) Structural model one-day ahead forecasts.

(b) GPR three-year ahead forecasts.

Figure 4: Visualization of the forecast experiment. The test data is plotted together with the h-step ahead forecasts and 95% confidence intervals for the structural model and Gaussian process regression (GPR).

Figure 4 shows the forecasts alongside the test data for the structural model and the GPR at the horizons where they perform best: 1D and 3Y, respectively. The test data (grey crosses) are plotted together with the h-step ahead forecasts (pink line) and

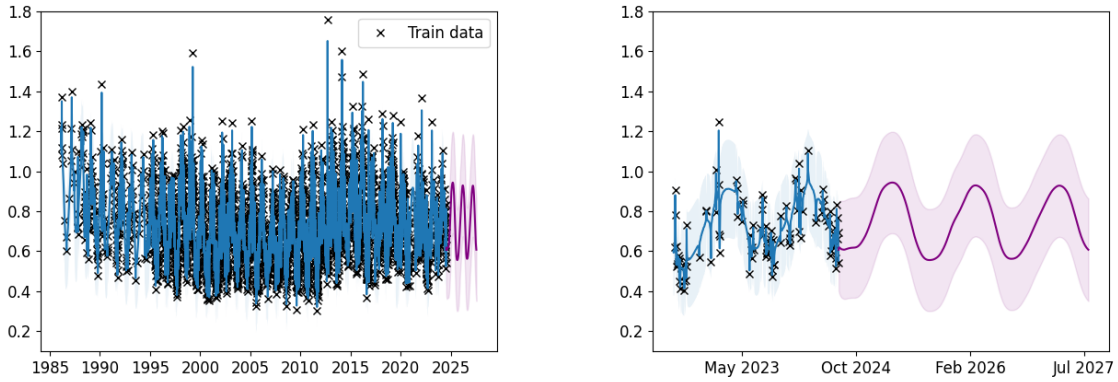
corresponding 95% confidence intervals (shaded area). The difference in FRMSE observed in Table 1 is clearly visible when comparing panels (a) and (b). The one-day ahead structural model forecasts (1D) closely follow the measurements, the vast majority of which fall within the confidence bands. The forecasts from the GPR, corresponding to the three-year horizon, are seen to be more stable. This pattern is intuitive: the most recent measurements typically have greater predictive power for nearby future observations than for those further ahead, hence it is often appropriate to assign more equal weight to the available data in long-term forecasts.

This principle also helps explain the results in Table 1. The extended kernel regression and structural models excel in short-term forecasting, which requires placing greater emphasis on recent observations. In contrast, models such as the trending seasonal model, GPR, DLS and DES (based on their estimated parameters) distribute weight more evenly across the available data, which helps maintain a stable performance across varying forecasting horizons.

4.2 Out-of-sample forecasts

Figure 5 shows the out-of-sample results for the atmospheric ethane burden above the Jungfraujoch station based on the structural model from Section 3.4. All forecasts are made at July 19$^{\text{th}}$, 2024, which is the end date of our sample ($t = n$). The forecasting horizon h ranges from one day up to three years, which gives a total of 1095 forecasts,

the last of which is for July 18$^{\text{th}}$, 2027. Panel (a) plots the observations (black crosses) as well as as the fit of the structural model (blue line) together with the forecasts (pink line) and 95% confidence intervals (shaded area). Panel (b) zooms in on the final year of the in-sample period and the forecast period. The observations (black dots) are shown together with the signal and the forecasts (pink line) as well as 95% confidence intervals (shaded area). In addition to a slight upward trend, we observe that the forecasts are mainly driven by the seasonal component.



(a) Data, fit and forecasts

(b) Final fit and forecasts

Figure 5: Out-of-sample forecasts and 95% confidence intervals for the atmospheric ethane burden above the Jungfraujoch station based on the Gaussian process regression model from Section 3.6. All forecasts are made at July 19$^{\text{th}}$, 2024, which is the end date of our sample ($t = n$). The forecasting horizon h ranges from one day up to three years, which gives a total of 1095 forecasts, the last of which is for July 18$^{\text{th}}$, 2027.

5 Conclusion

In this paper, we employed multiple approaches to forecast the atmospheric ethane burden above the Jungfraujoch measurement station in the Swiss Alps. The data is characterized by a large fraction of missing observations of around 76% and a strong seasonal pattern. We therefore employed a selection of time series forecasting methods which can handle these properties. In particular, we used five different approaches given by discounted least squares, damped exponential smoothing, a structural time series model, local linear kernel regression and Gaussian process regression.

Out-of-sample forecasts have been provided for the next three years. To investigate to what extent the different approaches are able to provide reliable ethane forecasts, we compared the forecasting ability of these approaches against a set of benchmarks in a forecasting experiment. The best performing benchmark was a deterministic trending seasonal model that has been used for the analysis of past trends in various time series of atmospheric ethane. Our main findings are that for the one-day forecasting horizon, the structural model provides the best forecasting results, while for longer horizons, the Gaussian process regression and damped exponential smoothing perform best. Overall, we conclude that the seasonal component remains stable over time and dominates the developments in atmospheric ethane, enabling reliable forecasting performance, even for longer horizons. This is reflected in our out-of-sample forecasts, which are mainly driven

by the seasonal component, in addition to predicting a slow-moving upward trend above the Jungfraujoch.

Forecasts of atmospheric ethane are important to monitor air quality and tropospheric pollution. In addition, they can be used to understand developments of methane released by the oil and gas sector. To the best of our knowledge, the analysis in this paper has been the first attempt in the literature in understanding potential future developments of atmospheric ethane. The conducted research therefore lends itself to a wider application of the proposed tools to other time series of atmospheric ethane, which could provide a more global view on future ethane developments.

References

Aikin, A. C., Herman, J. R., Maier, E. J., and McQuillan, C. J. (1982). Atmospheric chemistry of ethane and ethylene. *Journal of Geophysical Research*, 87(C4):3105–3118.

Anderson, B. D. and Moore, J. B. (2005). *Optimal filtering*. Courier Corporation.

Angelbratt, J., Mellqvist, J., Simpson, D., Jonson, J. E., Blumenstock, T., Borsdorff, T., Duchatelet, P., Forster, F., Hase, F., Mahieu, E., De Mazière, M., Notholt, J., Petersen, A. K., Raffalski, U., Servais, C., Sussmann, R., Warneke, T., and Vigouroux, C. (2011). Carbon monoxide (co) and ethane (c_2h_6) trends from ground-based solar ftir measurements at six european stations, comparison and sensitivity analysis with the emep model. *Atmospheric Chemistry and Physics*, 11(17):9253–9269.

Bernardi, M. and Catania, L. (2018). The model confidence set package for r. *International Journal of Computational Economics and Econometrics*, 8:144.

Brown, R. G. (1963). *Smoothing, forecasting and prediction of discrete time series*. Englewood Cliffs: Prentice Hall.

Cai, Z. (2007). Trending time-varying coefficient time series models with serially correlated errors. *Journal of Econometrics*, 136(1):163–188.

Chen, R., Yang, L., and Hafner, C. (2004). Nonparametric multistep-ahead prediction in time series analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66(3):669–686.

Chen, X. B., Gao, J., Li, D., and Silvapulle, P. (2018). Nonparametric estimation and forecasting for time-varying coefficient realized volatility models. *Journal of Business & Economic Statistics*, 36(1):88–100.

Collins, W. J., Derwent, R. G., Johnson, C. E., and Stevenson, D. S. (2002). The oxidation of organic compounds in the troposphere and their global warming potentials. *Climatic Change*, 52:453–479.

Diebold, F. and Rudebusch, G. (2022). On the evolution of us temperature dynamics. *Chudik, A., Hsiao, C. and Timmermann, A. (Ed.) Essays in Honor of M. Hashem Pesaran: Prediction and Macro Modeling (Advances in Econometrics)*, 43A:9–28.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263.

Durbin, J. and Koopman, S. J. (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford: Oxford University Press.

Fischer, E. V., Jacob, D. J., Yantosca, R. M., Sulprizio, M. P., Millet, D. B., Mao, J., Paulot, F., Singh, H. B., Roiger, A., Ries, L., Talbot, R. W., Dzepina, K., and Pandey Deolal, S. (2014). Atmospheric chemistry of ethane and ethylene. *Atmospheric Chemistry and Physics*, 14(5):2679–2698.

Franco, B., Bader, W., Toon, G. C., Bray, C., Perrin, A., Fischer, E. V., Sudo, K., Boone, C. D., Bovya, B., Lejeune, B., Servais, C., and Mahieu, E. (2015). Retrieval of ethane from ground-based FTIR solar spectra using improved spectroscopy: Recent burden increase above Jungfraujoch. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 160:36–49.

Franco, B., Mahieu, E., Emmons, L. K., Tzompa-Sosa, Z. A., Fischer, E. V., Sudo, K., Bovy, B., Conway, S., Griffin, D., Hannigan, J. W., Strong, K., and Walker, K. A. (2016). Evaluating ethane and methane emissions associated with the development of oil and natural gas extraction in North America. *Environmental Research Letters*, 11(4):44010.

Friedrich, M., Beutner, E., Reuvers, H., Smeekes, S., Urbain, J.-P., Bader, W., Franco, B., Lejeune, B., and Mahieu, E. (2020a). A statistical analysis of time trends in atmospheric ethane. *Climatic change*, 162:105–125.

Friedrich, M. and Lin, Y. (2024). Sieve bootstrap inference for linear time-varying coefficient models. *Journal of Econometrics*, 239(1):105345. Climate Econometrics.

Friedrich, M., Smeekes, S., and Urbain, J.-P. (2020b). Autoregressive wild bootstrap inference for nonparametric trends. *Journal of Econometrics*, 214(1):81–109.

Gardner, E. S. and McKenzie, E. (1985). Forecasting trends in time series. *Management science*, 31(10):1237–1246.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, 79(2):453–497.

Harvey, A. C. (1990). Forecasting, structural time series models and the kalman filter.

Harvey, A. C. and Phillips, G. D. (1979). Maximum likelihood estimation of regression models with autoregressive-moving average disturbances. *Biometrika*, 66(1):49–58.

Helmig, D., Rossabi, S., Hueber, J., Tans, P., Montzka, S. A., Masarie, K., Thoning, K., Plass-Duelmer, C., Claude, A., Carpenter, L. J., Lewis, A. C., Punjabi, S., Reimann, S., Vollmer, M. K., Steinbrecher, R., Hannigan, J. W., Emmons, L. K., Mahieu, E., Franco, B., Smale, D., and Pozzer, A. (2016). Reversal of global atmospheric ethane and propane trends largely due to US oil and natural gas production. *Nature Geoscience*, 9:490.

Holt, C. (1957). Forecasting seasonals and trends by exponentially weighted averages (onr memorandum no. 52). *Carnegie Institute of Technology, Pittsburgh USA*, 10.

Hyndman, R. J. and Athanasopoulos, G. (2021). Forecasting: principles and practice.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems.

Li, M., Pozzer, A., Lelieveld, J., and Williams, J. (2022). Northern hemispheric atmospheric ethane trends in the upper troposphere and lower stratosphere (2006–2016) with reference to methane and propane. *Earth System Science Data*, 14(9):4351–4364.

Lutsch, E., Strong, K., Jones, D., Blumenstock, T., Conway, S., Fisher, J., Hannigan, J., Hase, F., Kasai, Y., Mahieu, E., Makarova, M., Morino, I., Nagahama, T., Notholt, J., Ortega, I., Palm, M., Poberovskii, A., Sussmann, R., and Warneke, T. (2020). Detection and attribution of wildfire pollution in the arctic and northern midlatitudes using a network of fourier-transform infrared spectrometers and geos-chem. *Atmospheric Chemistry and Physics*, 20(20):12813–2851.

Maddanu, F. and Proietti, T. (2023). Trends in atmospheric ethane. *Climatic Change*, 176(5):1–23.

Ortega, I., Gaubert, B., Hannigan, J. W., Brasseur, G., Worden, H. M., Blumenstock, T., Fu, H., Hase, F., Jeseck, P., Jones, N., Liu, C., Mahieu, E., Morino, I., Murata, I., Notholt, J., Palm, M., RÃ¶hling, A., TÃ©, Y., Strong, K., Sun, Y., and Yamanouchi, S. (2023). Anomalies of o3, co, c2h2, h2co, and c2h6 detected with multiple ground-based fourier-transform infrared spectrometers and assessed with model simulation in 2020: Covid-19 lockdowns versus natural variability. *Elementa: Science of the Anthropocene*, 11(1):00015.

Robinson, P. M. (1989). *Nonparametric Estimation of Time-Varying Parameters*, pages 253–264. Springer Berlin Heidelberg.

Rudolf, J. (1995). The tropospheric distribution and budget of ethane. *Journal of Geophysical Research*, 100(D6):11369.

Schaefer, H. (2019). On the causes and consequences of recent trends in atmospheric methane. *Current Climate Change Reports*, 5(4):259–274.

Sun, Y., Yin, H., Liu, C., Mahieu, E., Notholt, J., Té, Y., Lu, X., Palm, M., Wang, W., Shan, C., Hu, Q., Qin, M., Tian, Y., and Zheng, B. (2021). The reduction in c_2h_6 from 2015 to 2020 over hefei, eastern china, points to air quality improvement in china. *Atmospheric Chemistry and Physics*, 21(15):11759–11779.

Szopa, S., Naik, V., Adhikary, B., Artaxo, P., Berntsen, T., Collins, W., Fuzzi, S., Gallardo, L., Kiendler-Scharr, A., Klimont, Z., Liao, H., Unger, N., and Zanis, P. (2021). Short-lived climate forcers. climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA*, pages 817–922.

Vinciguerra, T., Yao, S., Dadzie, J., Chittams, A., Deskins, T., Ehrman, S., and Dickerson, R. R. (2015). Regional air quality impacts of hydraulic fracturing and shale natural gas activity: evidence from ambient VOC observations. *Atmospheric Environment*, 110:144–150.

Visschedijk, A. J. H., Denier van der Gon, H. A. C., Doornenbal, H. C., and Cremonese, L. (2018). Methane and ethane emission scenarios for potential shale gas production in europe. *Advances in Geosciences*, 45:125–131.

Xiao, Y., Logan, J. A., Jacob, D. J., Hudman, R. C., Yantosca, R., and Blake, D. R. (2008). Global budget of ethane and regional constraints on U.S. sources. *Journal of Geophysical Research*, 113:D21306.

Young, P. C., Ng, C. N., Lane, K., and Parker, D. (1991). Recursive forecasting, smoothing and seasonal adjustment of non-stationary environmental data. *Journal of Forecasting*, 10(1-2):57–89.

A Implementation details

A.1 Forecasting with the structural model

The SSM formulation allows for estimating the states via the KF. Let $Y_t = \{y_1, \ldots, y_t\}$ denote the data available for conditioning at time t. The KF recursively computes for $t = 1, \ldots, n$ the filtering mean and variance

$$\mathrm{E}[\alpha_t|Y_t] \qquad \text{and} \qquad \mathrm{var}[\alpha_t|Y_t].$$

These can be used to perform h-step forecasting for $h > 0$ as follows. First note that unfolding h times the recursion for the state in (3.6) yields

$$\begin{aligned}\alpha_{t+h} &= T\alpha_{t+h-1} + c + \eta_{t+h-1} = T(T\alpha_{t+h-2} + c + \eta_{t+h-2}) + c + \eta_{t+h-1} = \dots \\ &= T^h\alpha_t + \sum_{j=0}^{h-1} T^j \left(c + \eta_{t+h-(j+1)}\right).\end{aligned}$$

From this expression we obtain the h-step forecasts of the states,

$$\begin{aligned}\mathrm{E}[\alpha_{t+h}|Y_t] &= \mathrm{E}\left[T^h\alpha_t + \sum_{j=0}^{h-1} T^j \left(c + \eta_{t+h-(j+1)}\right)\middle| Y_t\right] = T^h\,\mathrm{E}[\alpha_t|Y_t] + \sum_{j=0}^{h-1} T^j \left(c + \mathrm{E}[\eta_{t+h-(j+1)}|Y_t]\right) \\ &= T^h\,\mathrm{E}[\alpha_t|Y_t] + \sum_{j=0}^{h-1} T^j c,\end{aligned}$$

as $\mathrm{E}[\eta_{t+h-(j+1)}|Y_t] = \mathrm{E}[\eta_{t+h-(j+1)}] = 0$ for $j = 0, \dots, h-1$ with $h > 0$, since the state noise terms η_k are independent of Y_t for $k \geq t$. The h-step forecast variance is

$$\begin{aligned}\mathrm{var}[\alpha_{t+h}|Y_t] &= \mathrm{var}\left[T^h\alpha_t + \sum_{j=0}^{h-1} T^j\eta_{t+h-(j+1)}\middle| Y_t\right] = \mathrm{var}\left[T^h\alpha_t\middle|Y_t\right] + \mathrm{var}[\sum_{j=0}^{h-1} T^j\eta_{t+h-(j+1)}] \\ &= T^h\,\mathrm{var}[\alpha_t|Y_t]\left(T^h\right)' + \sum_{j=0}^{h-1} T^jQ\left(T^j\right)',\end{aligned}$$

where the second equality follows as η_k is independent of α_t and Y_t for $k \geq 0$, and the final expression follows from serial independence of the state errors.

Our main focus is forecasting the signal $f_t \equiv Z_t\alpha_t$ and the corresponding noisy measurements $y_t = f_t + \varepsilon_t$. The relevant quantities can be expressed in terms of those that were obtained for the states. The forecasting means are

$$\mathrm{E}[y_{t+h}|Y_t] = \mathrm{E}[f_{t+h} + \varepsilon_{t+h}|Y_t] = \mathrm{E}[f_{t+h}|Y_t] = Z_{t+h}\,\mathrm{E}[\alpha_{t+h}|Y_t],$$

since ε_{t+h} is independent of Y_t for $h > 0$. From similar reasoning, it follows that the forecasting variances are

$$\operatorname{var}[y_{t+h}|Y_t] = \operatorname{var}[f_{t+h} + \varepsilon_{t+h}|Y_t] = \operatorname{var}[f_{t+h}|Y_t] + \operatorname{var}[\varepsilon_{t+h}] = Z_{t+h} \operatorname{var}[\alpha_{t+h}|Y_t] Z'_{t+h} + H.$$

A.2 SSM estimation based on diffuse log likelihood

In order to apply the structural time series model, the parameter vector θ defined in (3.5) must be estimated using the available measurements. The standard approach for linear Gaussian SSMs is to use maximum likelihood estimation (Durbin and Koopman, 2012, Ch. 7). We consider the following decomposition of the likelihood,

$$\mathcal{L}(\theta) \equiv p(y_1, y_2, \ldots, y_n) = p(y_1) \prod_{t=1}^{n-1} p(y_{t+1}|Y_t),$$

which follows from Bayes' formula, with p denoting a probability density function for the corresponding variables. The measurements $y_1, y_2, \ldots, y_n$ are jointly normal because of the assumed independence between α_1, ε_t, and η_k for $k, t = 1, \ldots, n$. In this case, the conditional densities $p(y_{t+1}|Y_t)$ are also normal (Anderson and Moore, 2005, Example 3.2), hence they are completely determined by the one-period forecasting means $\mathrm{E}[y_{t+1}|Y_t]$ and variances $\operatorname{var}[y_{t+1}|Y_t]$. As discussed in the previous section, for any value of θ these means and variances can be computed by the KF, which allows for computing the likelihood $\mathcal{L}(\theta)$. However, since diffuse initialization is used for the elements μ_1 and δ_1, it is preferable to use a diffuse analogue of the likelihood (Durbin and Koopman, 2012, Ch. 7.2.2), which we

define by

$$\mathcal{L}_d(\theta) = \prod_{t=2}^{n-1} p(y_{t+1}|Y_t). \tag{A.1}$$

The above *diffuse likelihood* omits the first two terms of the likelihood $\mathcal{L}(\theta)$, which is because for our model these terms diverge when $\kappa \to \infty$. Loosely speaking, we require two observations, y_1 and y_2, to learn something about the direction of the trend as

$$\delta_1 = \mu_2 - \mu_1 = y_2 - y_1 + (s_1 + \varepsilon_1 - s_2 - \varepsilon_2).$$

In practice, we therefore perform maximum likelihood estimation by numerically maximizing the average diffuse log likelihood, $\log\left\{\mathcal{L}_d(\theta)\right\}/n$, over feasible values of θ.

A.3 Local linear kernel estimation

The time-varying coefficient model can be considered an extension of the non-parametric trend model by allowing for the inclusion of exogenous regressors. The model is given by:

$$y_t = \theta_t' x_t + \varepsilon_t, \quad t = 1, \ldots, n, \tag{A.2}$$

where $\theta_t = \theta(t/n)$ and x_t are defined in Section 3.5 and ε_t denotes the error term. We define that all functions are in $\mathcal{C}^3[0,1]$ which is a common smoothness assumption required for non-parametric estimation. Then, for t/n in an ρ-neighborhood around $\tau \in (0,1]$, each coefficient curve can be locally approximated using a first-order Taylor approximation. For the trend function $g(\cdot)$ this would yield $g(t/n) \approx g(\tau) + g^{(1)}(\tau)(t/n - \tau)$, with $g^{(1)}(\cdot)$

denoting the first derivative of $g(\cdot)$. Let us define $\tau_t := t/n$. Using the local approximation, we can reformulate model (A.2) as

$$y_t \approx \theta(\tau)'x_t + \theta^{(1)}(\tau)'x_t(\tau_t - \tau) + \varepsilon_t =: \tilde{x}_t(\tau)'\psi(\tau) + \varepsilon_t, \tag{A.3}$$

where $\theta^{(1)}(\tau) = \left(g^{(1)}(\tau), a_1^{(1)}(\tau), \ldots, a_S^{(1)}(\tau), b_1^{(1)}(\tau), \ldots, b_S^{(1)}(\tau)\right)'$, $\psi(\tau) = \left(\theta(\tau)', \theta^{(1)}(\tau)'\right)'$, and $\tilde{x}_t(\tau) = (x_t', x_t'(\tau_t - \tau))'$. This local approximation ensures that we can estimate the coefficient curves by the local linear estimator, which is obtained by minimizing the following weighted sum of squares:

$$\widehat{\psi}(\tau) = \begin{pmatrix} \widehat{\theta}(\tau) \\ \widehat{\theta}^{(1)}(\tau) \end{pmatrix} = \underset{\psi(\tau)}{\operatorname{argmin}} \sum_{t \in D_t}^{n} \left(y_t - \tilde{x}_t(\tau)'\psi(\tau)\right)^2 K\left(\frac{\tau_t - \tau}{\rho}\right), \tag{A.4}$$

where D_t denotes the set of time indices t for which the measurements y_t are complete. Furthermore, $K(\cdot)$ is a kernel function with bandwidth $\rho > 0$, which embodies the concept that model (A.3) is a local approximation. The closed-form expression of the solution to the minimization problem in (A.4) is given by

$$\widehat{\psi}(\tau) = \begin{pmatrix} \widehat{\theta}(\tau) \\ \widehat{\theta}^{(1)}(\tau) \end{pmatrix} = \begin{pmatrix} S_{n,0}(\tau) & S_{n,1}'(\tau) \\ S_{n,1}(\tau) & S_{n,2}(\tau) \end{pmatrix}^{-1} \begin{pmatrix} T_{n,0}(\tau) \\ T_{n,1}(\tau) \end{pmatrix} =: S_n^{-1}(\tau)T_n(\tau), \tag{A.5}$$

where $\tau \in [0, 1]$, and for $k = 0, 1, 2$ we have:

$$\begin{aligned} S_{n,k}(\tau) &= \frac{1}{n\rho} \sum_{t \in D_t}^{n} x_t x_t' \left(\tau_t - \tau\right)^k K\left(\frac{\tau_t - \tau}{\rho}\right), \\ T_{n,k}(\tau) &= \frac{1}{n\rho} \sum_{t \in D_t}^{n} x_t \left(\tau_t - \tau\right)^k K\left(\frac{\tau_t - \tau}{\rho}\right) y_t. \end{aligned} \tag{A.6}$$

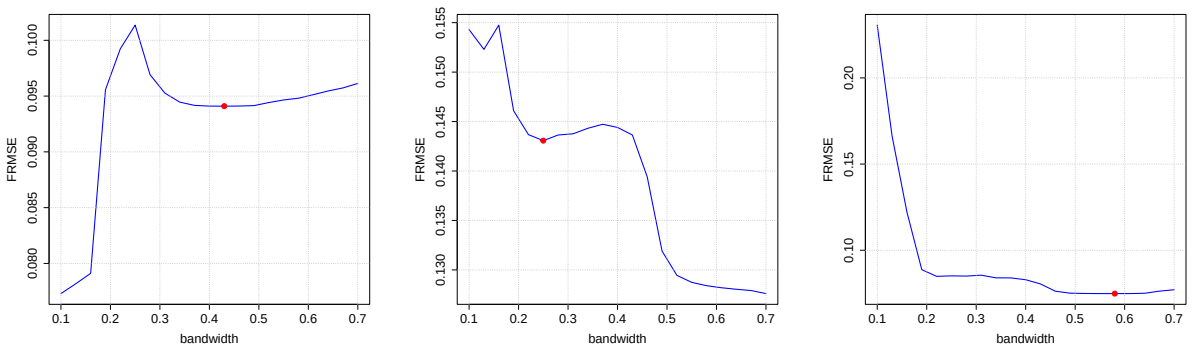
A.4 Bandwidth selection

As mentioned in Section 3.5, the local linear estimator $\widehat{\theta}(\tau)$ is obtained by fitting a locally weighted linear regression to the data in an ρ-neighborhood around τ. The parameter ρ, called the bandwidth, controls the width of the neighborhood. It is crucial to select a suitable bandwidth ρ as it controls the model complexity and there is a bias-variance tradeoff. We consider block cross-validation (BCV) for time series forecasting where we divide the training sample into k blocks in time, such that each block contains the same number of complete measurements. Within each block, we divide the complete measurements into a training and test set based on a 90-10 split. Subsequently, for bandwidths ranging from 0.1 to 0.7 in increments of 0.03, we fit the model to the training data of a block and forecast each point in the test set of this block as an h-step ahead forecast. We then calculate the corresponding h-step ahead forecast errors $e_{t+h|t}$ for each bandwidth and compute the FRMSE.

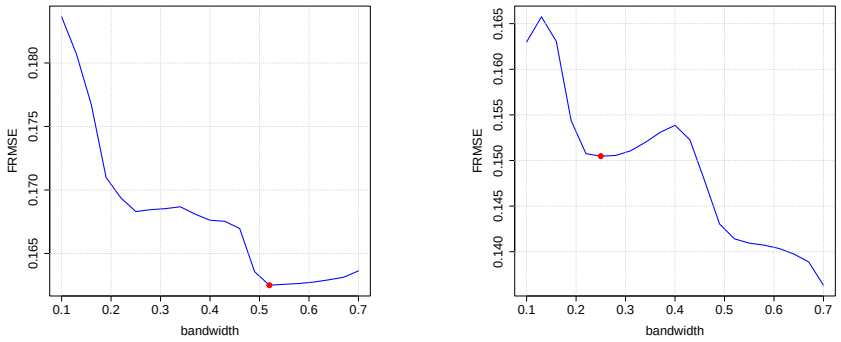
The objective is to find the optimal bandwidth that minimizes the FRMSE within a block. This gives us an optimal bandwidth per block, for each forecast horizon. For our forecasting experiment, we divide the data into $k = 5$ separate blocks. Moreover, since the multistep ahead estimator consists of a collection of one-step ahead forecasts, we obtain the optimal bandwidth for a forecasting horizon of $h = 1$. The small forecasting horizon is further motivated by the increased computational burden of the BCV method

when h is large.

As data-driven methods for bandwidth selection tend to select bandwidths near the boundaries of the grid, we visually inspect the FRMSE as a function of the bandwidth to identify local minima. Specifically, for each of the $k = 5$ blocks, the block optimal bandwidth is given by the local minimum resulting in the smallest one-step ahead FRMSE. Subsequently, to combine the results from the different blocks, $\widehat{\rho}$ is obtained by taking the average of the block optimal bandwidths. In Figure 6 and Figure 7 we visualize this procedure for the base model consisting of a trend and seasonal terms, and the extended model including lags of the dependent variable, respectively. As can be observed from Figure 7, block 4 does not contain local minima for the extended model, such that we do not identify a block optimal bandwidth. Therefore, for the extended model, $\widehat{\rho}$ is obtained by taking the average block optimal bandwidth over the remaining blocks. Table 2 complements the figures by presenting the optimal bandwidth per block and the resulting selected bandwidth $\widehat{\rho}$ for both models.

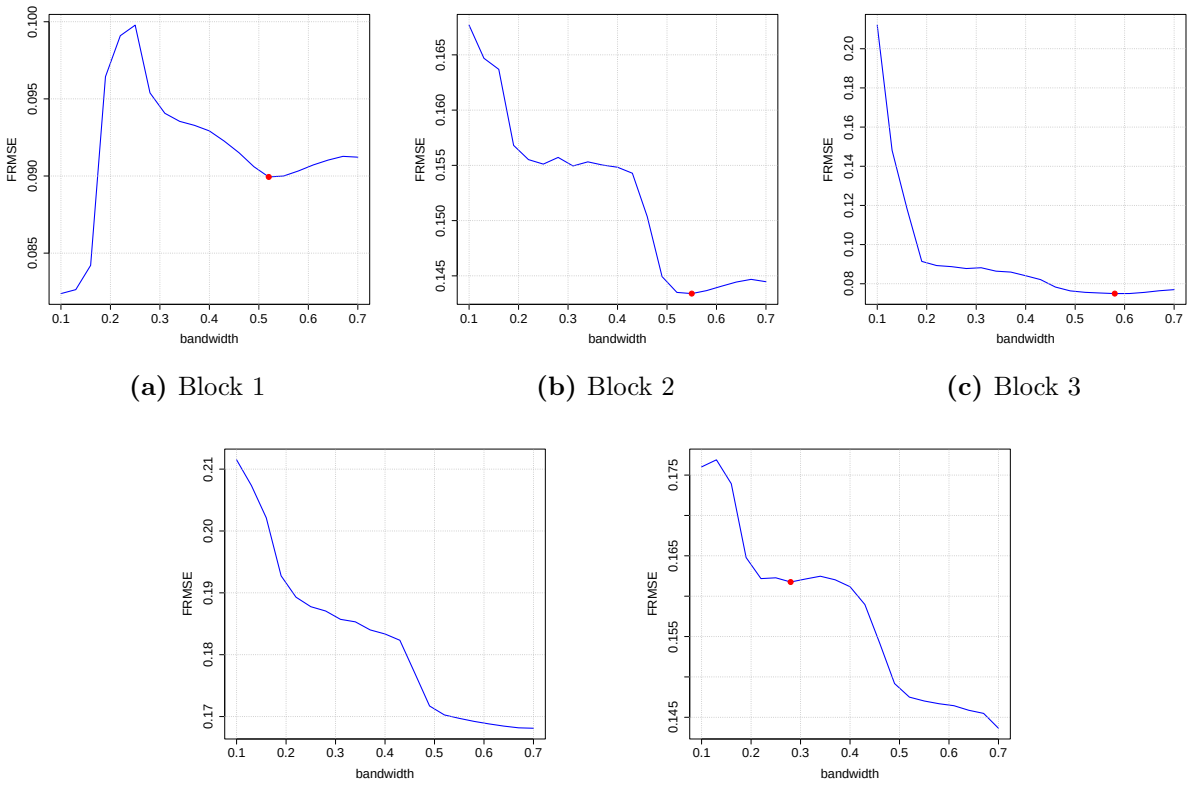


(a) Block 1 **(b)** Block 2 **(c)** Block 3



(d) Block 4 **(e)** Block 5

Figure 6: one-step ahead FRMSE per block as a function of the bandwidth for the base model consisting of a trend and seasonal terms. The red dot indicates the selected local minimum, which corresponds to the optimal block bandwidth.



(a) Block 1 **(b)** Block 2 **(c)** Block 3

(d) Block 4 **(e)** Block 5

Figure 7: one-step ahead FRMSE per block as a function of the bandwidth for the extended model including lags of the dependent variable. The red dot indicates the selected local minimum, which corresponds to the optimal block bandwidth.

	Block 1	Block 2	Block 3	Block 4	Block 5	$\widehat{\rho}$
Base model	0.43	0.25	0.58	0.52	0.25	0.41
Extended model	0.52	0.55	0.58	-	0.28	0.48

Table 2: Optimal bandwidth per block and the selected bandwidth $\widehat{\rho}$ for the base model consisting of a trend and seasonal terms, and the extended model including lags of the dependent variable. Note that $\widehat{\rho}$ is rounded up to two decimal places.