Methods for missing time-series data and large spatial data

Rachael Duncan, MSci (Hons), MSc

School of Mathematics and Statistics

Lancaster University

A thesis submitted for the degree of

Doctor of Philosophy

December, 2023

Methods for missing time-series data and large spatial data

Rachael Duncan, MSci (Hons), MSc.

School of Mathematics and Statistics, Lancaster University

A thesis submitted for the degree of *Doctor of Philosophy*. December, 2023.

Abstract

Performing accurate statistical inference requires high-quality datasets. However, real-world datasets often contain missing variables of varying degrees both spatially and temporally. Alternatively, modelled datasets can provide a complete dataset, but these are often biased. This thesis derives a simplified approach to the skew Kalman filter that tackles the computational issues present in the existing skew Kalman filter by using a secondary dataset to estimate the skewness parameter. In application, this thesis implements the skew Kalman filter using surface-level ozone to bias-correct the modelled ozone data and use the bias-corrected data to infill missing data in the observed dataset. Further, this thesis explores working with large spatial datasets. When carrying out spatial inference, using all the possible data available allows for more accurate inference. However, spatial models such as Gaussian processes scale cubically with the number of data points and thus quickly become computationally infeasible for moderate to large datasets. Divide-and-conquer methods allow data to be split into subsets and inference is carried out on each subset before recombining. While well documented in the independent setting, these methods are less popular in the spatial setting. This thesis evaluates the performance of divide-and-conquer methods in the spatial setting to achieve approximate results compared to carrying out inference on the full dataset. Finally, this is demonstrated using USA temperature data.

Acknowledgements

Firstly, I would like to thank my supervisors Professor Chris Nemeth and Dr Paul Young for their guidance and support throughout my PhD. I would also like to thank Professor Theodore Kypraios and Dr Caroline Euan for examining my thesis, your comments and feedback improved my thesis, and also for an honestly enjoyable viva experience.

A PhD at Lancaster was made so much more enjoyable by the people met along the way. I would like to thank all those in the DSNE office for their support whether to bounce ideas off or enjoy. Especially to Kate Wright and Sarah Jones, our Friday morning writing groups were always a welcome end to the week.

Finally, I would like to thank my gran for her unwavering support in my academic endeavours, and my partner Ben for helping make my time outside the PhD so enjoyable and providing much needed balance to my life with so much joy and support.

Declaration

I declare that the work presented in this thesis is, to the best of my knowledge and belief, original and my own work. The material has not been submitted, either in whole or in part, for a degree at this, or any other university. This thesis does not exceed the maximum permitted word length of 80,000 words including appendices and footnotes, but excluding the bibliography.

Rachael Duncan

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Our understanding of the world around is driven by the data available to us. Environmental monitoring is an important aspect of understanding the environment around us and the impact it has. In 2016 as much as 24% of worldwide deaths were related to environmental factors. Changes occur in the environment, not only in natural cycles but from anthropogenic-based impacts as well and we observe these changes through environmental monitoring (Pruss-Ustun et al. 2018, Forouzanfar et al. 2016).

Networks of monitoring stations exist across the globe monitoring environmental factors, including monitoring temperature for extreme events such as heat waves and cold snaps, rainfall data and water levels for flood risk and air quality for high pollution episodes that impact human health and the environment (Lovett et al. 2007). Effects from environmental changes such as pollution on humans can be slow or fast acting and occur at all spatial scales (Cohen et al. 2017). Pairing statistical methods with environmental monitoring is necessary as it is impossible to characterise environmental changes everywhere all the time. Statistical inference is widely used to infer or interpolate information from data, such as recovering missing data or estimating some environmental variable at unseen locations (Zhang et al. 2012), practically monitoring every street in a city or part of a river is impossible.

Missing data is important to address since it can be a barrier to understanding. Temporally or spatially incomplete data can result in less accurate inference and can introduce biases into the output from statistical models. In this thesis, we tackle two main problems, using examples from two important environmental fields. Many methods for infilling missing data exits ranging from simple to complex (Hartley & Hocking 1971, Little & Rubin 2002, and references therein). Simple methods, while straightforward, and easy to use can disrupt the structure of the data and introduce large errors in the analysis (Baraldi & Enders 2010, Donders et al. 2006). Although more sophisticated approaches are often more accurate, these methods can be computationally expensive or require additional covariate information, particularly if modelling the data is required. This thesis focusses on developing and implementing methods that are computationally efficient for working with environmental data.

1.1 Motivating Applications

The first application of interest is temporally missing data, for this we use the example of air pollution. The World Health Organization (2022) estimate that 4.2 million die prematurely from outdoor air pollution each year and that 99% of the world's population live in areas not meeting safe air quality guidelines. Therefore it is important to be able to identify primary sources that lead to poor air quality, quantify the impact of these sources on air quality, and reduce the threats of detrimental air quality on human health and the environment. To do this we require good-quality data to carry out inference. However, environmental datasets often contain missing data due to sensor faults, maintenance, and repair. Figure 1.1.1 shows the locations of background ozone monitoring stations in Germany from the European Environment Agency (https://www.eea.europa.eu/data-and-maps). The amount of missing data varies across the network, for example looking at the hourly

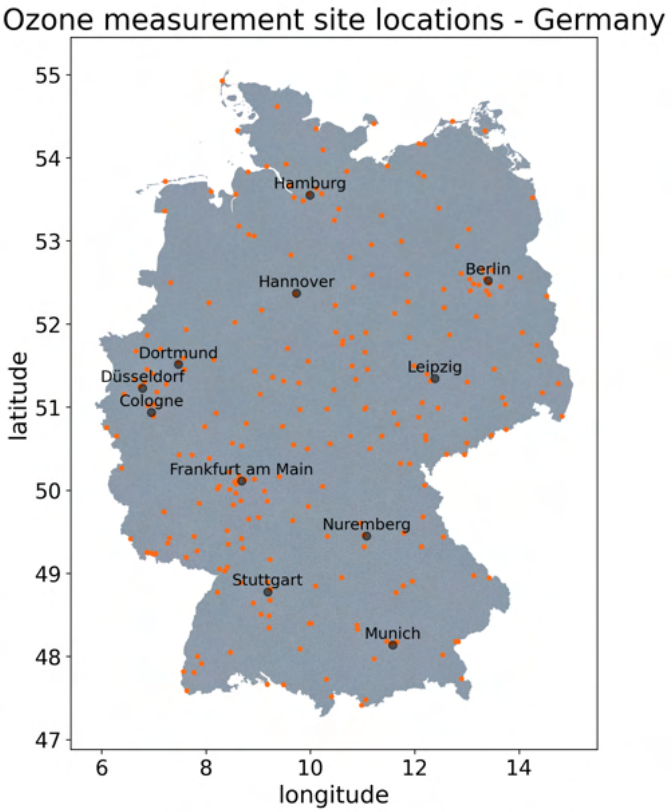


Figure 1.1.1: Map of ozone monitoring stations in Germany.

time series for a site in Hannover, in 2018 there were 538 missing data points, which is 6.14% of the data for that year. Over the same period, a site in Braunschweig was missing 373 data points, which is 4.25%.

While physically based models can be used to produce a temporally and spatially complete dataset these are often biased, particularly at local scales, undermining its use as a straight alternative to monitoring data. To address this problem, we look at infilling missing surface-level ozone data from monitoring stations using temporally complete biased data. To do this we develop a skew Kalman filter approach that allows us to conceptualise the bias as a skew between the reanalysis data and the monitoring data and correct for that bias with an estimated uncertainty for the infilled data.

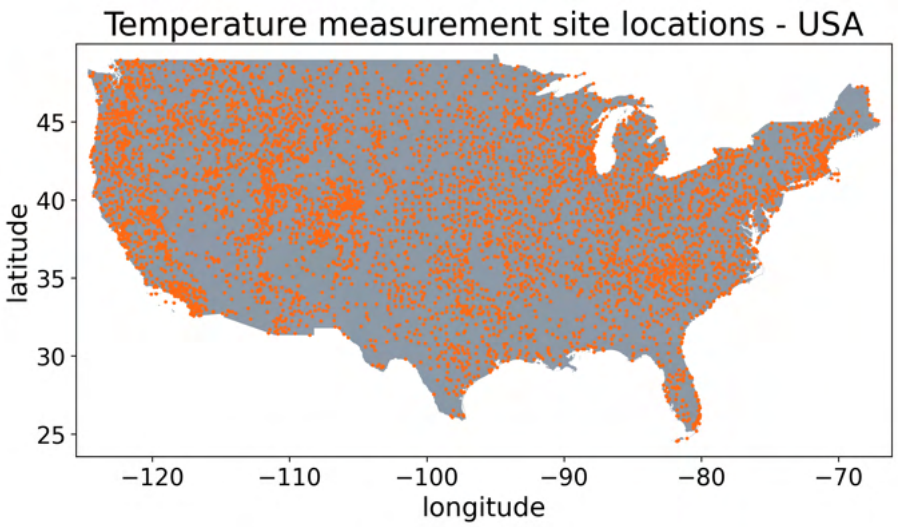


Figure 1.1.2: Map of temperature monitoring stations in the USA.

The second application considers spatially missing data; we use the example of temperature monitoring stations in the USA. Temperature monitoring is important for understanding and quantifying changes in the mean and extremes. Extreme heat events have been increasing in frequency, duration and intensity over recent decades, which has been linked to climate change (IPCC 2014), and which is responsible for over 600 deaths in the United States each year NCEH (2020). Figure 1.1.2 shows the location of measurement sites in the USA that had annual temperature averages in 2018, from 7196 locations only 5660 had an annual average temperature, which is 21.4% missing, leaving large areas unmonitored.

For this work, we consider approaches for dealing with large spatial datasets. Large datasets pose a challenge when carrying out inference due to limits in memory, ram, processing speeds and data storage. Divide and conquer approaches allow the data to be divided into subsets, and inference carried out on each subset before combining is a popular tool for handling large datasets. Various methods for combining the subsets through a divide-and-conquer approach exist in the independent setting, including consensus Monte Carlo (Scott et al. 2016), Gaussian Process barycentres (Mallasto & Feragen 2017), and SwISS (Sub-posteriors with Inflation, Scaling and

Shifting) (Vyner et al. 2022). However, their usefulness in the spatial setting has not been fully evaluated. Methods which allow the full dataset to be utilised should allow for more accurate model output across the domain.

1.2 Thesis Structure

This thesis focusses on approaches for incomplete data and the contributions of this these are as follows: the development of a new approach to the skew Kalman filter; the demonstration of the skew Kalman filter as a tool for bias correction and missing data infilling; a demonstration of how existing divide and conquer approaches perform in the spatial setting. The structure of the thesis is given below

- Chapter 2 presents a literature review of missing data approaches in environmental monitoring.
- Chapter 3 introduces the background material for Bayesian optimal filtering and derives the traditional Kalman filter.
- Chapter 4 introduces an extension to the traditional Kalman filter that allows for skewness to be present in the observation noise. A new approach to the skew Kalman filter is developed which simplifies how parameters in the skew Kalman filter are estimated. This work is demonstrated using a simulation study.
- Chapter 5 implements the skew Kalman filter as a tool for bias correction and infilling missing data. Using daily mean surface level ozone data, the skew Kalman filter is used to bias correct reanalysis data. By simulating random missingness and periods of consecutive missing, the performance of the skew Kalman filter as a tool for infilling missing data is evaluated before being demonstrated on a real-world example.

- Chapter 6 evaluates the performance of divide and conquer approaches when fitting a Gaussian process model to spatial data. The combining strategies considered are consensus MCMC, SwISS and Gaussian process barycentres. First using a smaller simulated dataset, the methods are evaluated on how well they capture the model fit when using the full data, the runtime of each of the methods and how estimating values at unseen locations compares to using full data to estimate the values. This work is demonstrated on a larger dataset using USA temperature data, again evaluating runtime and how well the model fits compared to using the full data.

- Chapter 7 concludes the thesis, reviewing the main results of the thesis and discussing the avenues for future work.

Chapter 2

Missing Data in Environmental Datasets: A Review

Missing data is a pervasive problem in environmental research. This may be the result of insufficient sampling, errors in the measurements, or faults in data acquisition. Missing data results in discontinuities in the time series which pose significant obstacles for time-series prediction schemes which often require continuous data. The absence of data could cause bias in the statistical inference, leading to invalid conclusions (Zhang & Thorburn 2022, Noor et al. 2014, Baraldi & Enders 2010, Plaia & Bondì 2006, Donders et al. 2006, Gnauck & Luther 2005, Junninen et al. 2004). Thus, we need a method to infill missing values.

Missing data is an important and well studied problem in statistics. As such, statisticians have a well-developed topology to describe different patterns and mechanisms important in characterising missing data. In this chapter we will discuss the different types of missing data and, the missing data patterns that arise. We describe some of the well established traditional statistical methods for infilling missing data. We will then give overview of these traditional methods and their suitable applications. More recent methods try to address some of the issues present in the traditional methods and we give a brief overview of popular sophisticated

methods used to infill missing data. Finally, we give our conclusions.

2.1 Types of Missing Data

The choice of the appropriate method for handling missing data depends on the pattern of missing data and the mechanism for why it is missing. This gives rise to two common classifications for missing data types, as outlined in the sections below.

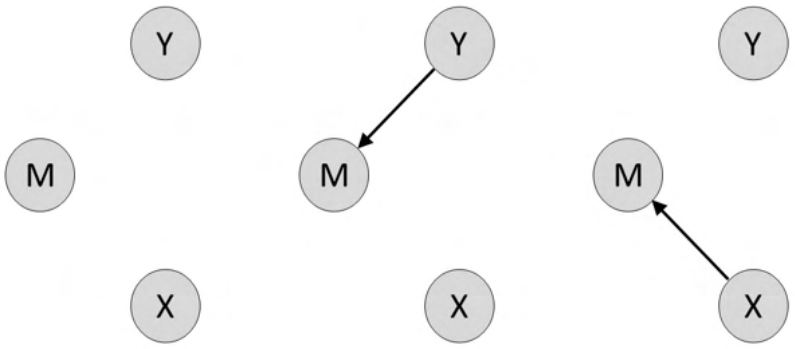
(1) Classification based on a missing mechanism: First proposed by Rubin (1976), the standard classification of missing data mechanism considers missing data as (a) Missing Completely at Random (MCAR), (b) Missing at Random (MAR), and (c) Missing Not at Random. The relationship between missingness (no data value is stored for the variable in an observation), completely observed variables, and the latent variables is described in terms of probability. The specific dependencies are shown in Figure 2.1.1, where Y represents the variables that are observed, X represents the latent variable and M represents missingness.

(a) Missing Completely at Random: Shown in Figure 2.1.1a, MCAR is the simplest of the three mechanisms and states that the probability of missingness is independent of observed and latent variables. The probability of missingness is given by

$$P(M|Y, X) = P(M) \tag{2.1.1}$$

(b) Missing at Random: Shown in Figure 2.1.1b, the probability of missingness is dependent on the observed variable, and is given by

$$P(M|Y, X) = P(M|Y) \tag{2.1.2}$$



(a) MCAR (b) MAR (c) MNAR

Figure 2.1.1: Dependencies for the different mechanisms for missing data

(c) Missing Not at Random: Shown in Figure 2.1.1c, MNAR is the most complex mechanism. Here, the probability of MNAR is dependent on the latent variable and is given by

$$P(M|Y, X) = P(M|X) \tag{2.1.3}$$

Of the above classifications, missing data is most often considered as MAR or MCAR. In monitoring systems, such as the ones used in environmental monitoring, MCAR is the most common mechanism as a monitoring site being down is an independent random event. MAR may result from successive missingness at several periods, as the missingness at a location is related to its adjacent locations. MNAR is caused by some latent factor such as limits in power or memory, MNAR is rarely studied as it is assumed these factors can be found and solved in advance (Du et al. 2020, Plaia & Bondì 2006).

(2) Classification based on missing patterns: Collected data can be seen as a large matrix, with the rows representing the number data and columns representing the different dimensions in the data. This allows us to characterise the missingness into different patterns of the matrix (Schafer & Graham 2002). The missing patterns can be divided into four categories (a) univariate pattern, (b) multivariate pattern, (c) monotone pattern, and (d) arbitrary pattern. These are defined as follows.

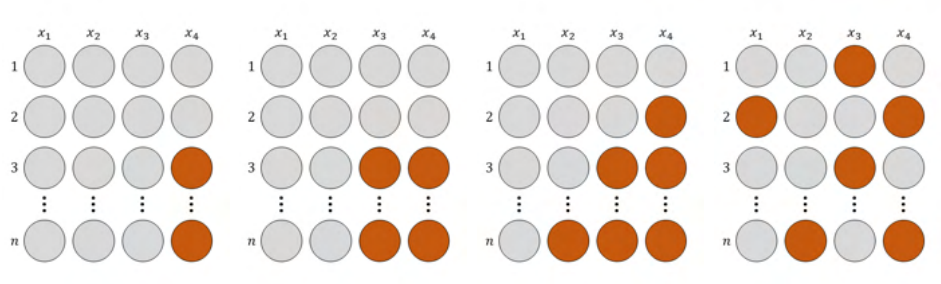
(a) Univariate missing pattern: Shown in Figure 2.1.2a missing data is limited to one dimension, here only the 4th column shows missingness.

(b) Multivariate missing pattern: Shown in Figure 2.1.2b missing data is distributed equally across multiple dimensions. Here columns 3 and 4 shows missingness, while the other columns are complete.

(c) Monotone missing pattern: Shown in Figure 2.1.2c, missing data is said to be monotone if it can be arranged to be ladder-like. This type of missing data is typically representative of a structured missing pattern.

(d) Arbitrary missing pattern: Shown in Figure 2.1.2d, this is the most common pattern of missingness and occurs when missing data are randomly distributed in different rows and columns.

The missing data mechanism MCAR would correspond to the arbitrary missing pattern and would be expected to be the missing data pattern for monitoring systems. However, univariate or multivariate missing best describes the missing pattern seen in monitoring systems (Du et al. 2020). This results as if a sensor is off for a period of time it results in successive missingness in 1 or more dimensions.



(a) univariate missing (b) multivariate missing (c) monotone missing (d) arbitrary missing

Figure 2.1.2: Patterns of missing data. Grey denotes normal values and orange denotes the missing values, where rows $1, 2, .., n$ represent the number or time-series of data and x_1, x_2, x_3, x_4 represent the different dimensions in the data. Adapted from Schafer & Graham (2002), Du et al. (2020).

2.2 Methods for Infilling Missing Data

2.2.1 Traditional Methods

Various methods for recovering missing data are used in the literature. These methods include case deletion, mean method, nearest neighbour, linear and cubic spline interpolation, regression based methods, expectation maximisation, K-nearest neighbours, and multiple imputation (Allison 2001, Meijering 2002, Enders 2022, and references therein). In this section we summarise these classical methods.

2.2.1.1 Casewise Deletion

The simplest approach for handling missing data is casewise deletion, also known as listwise deletion. All entries with missing data are removed or discarded when doing analysis, meaning estimating the missing value is unnecessary. Deletion can introduce bias in the analysis, especially when the missing data is MAR or MNAR, as opposed to MCAR where it is less likely to be an issue (Little & Rubin 2002). When working with time-series data, deletion can disrupt the data structure (Baraldi & Enders 2010, Donders et al. 2006). A fundamental issue with removing all missing

entries is that this will reduce the sample size which will be increasingly problematic with increasing missing data.

2.2.1.2 Mean Method

The mean method fills in missing values with the mean of the observed data. Due to its simplicity, it was one of the most commonly used methods in the analysis of data from early environmental monitoring systems (Junninen et al. 2004, Noor et al. 2014) and is considered a baseline method. Compared to casewise deletion, the sample size is preserved. While the sample mean statistic remains unchanged after imputation, the sample variance changes as follows

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}[a_i(y_i-\bar{y})^2 + (1-a_i)(\hat{y}_i-\bar{y})^2], \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2, \\ &= \frac{n_1-1}{n_1}s_1^2, \end{aligned} \tag{2.2.1}$$

where a_i is the indicator vector, when $a_i = 0$ data are observed and when $a_i = 1$ data as missing. n and n_1 are the number of data and observed data respectively, s_1^2 is the sample variance of the observed data and $\bar{y}$ is the sample mean. This results in a smaller estimated variance than the actual variance as the infilled results are too concentrated about the mean. Methods such as hierarchical mean method (Kiani & Saleem 2017) address the issues in the variance by diving the data into groups based on similarity and then applying the mean method to each group.

2.2.1.3 Nearest Neighbour and Linear Interpolation

Univariate nearest neighbour imputation is a simple scheme for infilling missing values. The endpoints of each gap are used as estimates for all the missing values

such that

$$\begin{aligned} y &= y_1 \qquad \text{if } x \leq x_1 + (x_2 - x_1)/2, \\ y &= y_2 \qquad \text{if } x > x_1 + (x_2 - x_1)/2, \end{aligned} \tag{2.2.2}$$

where y is the interpolant, x is the time point of the interpolant and (x_1, y_1) are the coordinates of the starting point in the gap, and (x_2, y_2) are the endpoint coordinates.

Linear interpolation fits a straight line between the end points of the gap. Missing values are calculated using the line equation

$$y = y_1 + k(x + x_1) \quad \text{where } k = \frac{(y_2 - y_1)}{x_2 - x_1};\ x_1 < x < x_2 \text{ and } y_1 < y < y_2. \tag{2.2.3}$$

2.2.1.4 Regression-Based Methods

Regression-based methods are based on estimated regression models between missing data and available data, as a predictor. Using a correlation matrix, several predictors of the variable with missing data is identified and the best predictors are selected and used as independent variables in a regression equation (Braak et al. 1994, Carvalho et al. 2011, Ohba et al. 2016, Liu et al. 2017, Mirzaei et al. 2022). The underlying model is given by

$$y = f(x) + \epsilon \tag{2.2.4}$$

where $x \in \mathbb{R}^d$ is the independent variable, $y \in \mathbb{R}$ is the variable containing missing information, and ϵ is a noise term given by $\epsilon \sim N(0, \sigma^2)$. The aim is to find a model, $g(\cdot)$, that approximates $f(\cdot)$, such that we can use the well-learned model to recover the missing information. In the case of linear regression this will be modelled using a linear function. However, as the relationship between variables will not always be linear and other regression-based methods can be used, including polynomial regression (Carvalho et al. 2011) and log-linear regression Braak et al. (1994).

Since the correlation between the independent variables and the missing variable cannot be known in advance, it takes time to find the optimised form of the regression model. Thus, the main issue facing regression-based methods is determining $g(\cdot)$. Also, the need that the independent variables must not contain missing information, can limit the applications suitable for this method.

2.2.1.5 Expectation-Maximisation

The expectation-maximisation algorithm is an iterative algorithm for performing maximum likelihood estimation in the presence of latent variables (Dempster et al. 1977, Qu et al. 2009), that can be used for infilling missing data (Ghahramani & Jordan 1995, Schneider 2001, Baraldi & Enders 2010, Zhang et al. 2015). Compared with the methods outlined above, it directly models incomplete data and generates estimated values using maximum likelihood estimation. The expectation-maximisation algorithm consists of two parts: the E-step, which calculates the conditional expectation of missing data based on the observation data and current model parameters, and the M-step, which uses the conditional expectation to infill the missing values. The model parameters are updated using maximum likelihood estimation and the algorithm iterates between the steps until the model parameters converge. The expectation-maximisation algorithm steps are as follows:

1. The E-step calculates the conditional expectation of missing data based on all observed values and current estimated model parameters

$$Q(\theta|\theta^i, y) = \int \log(\theta|Y) f(Y_{missing}|Y_{obs}, \theta^i) dY_{missing}. \tag{2.2.5}$$

   where $\log(\theta|Y)$ is the log likelihood function of the complete data and $f(Y_{missing}|Y_{obs}, \theta^i)$ is the predictive distribution of the missing data given θ.

2. The M-step obtains the next estimate of θ by maximising $Q(\theta|\theta^i)$,

$$\theta^{i+1} = \arg\max_{\theta} Q(\theta|\theta^i). \tag{2.2.6}$$

The EM algorithm converges when $\|\theta^{i+1} - \theta^i\| < \epsilon$, where ϵ is some predefined threshold.

The EM algorithm holds well for large-scale data. However, for practical applications, the convergence speed is limited by the selection of initial values. When processing severely missing data, computation slows and the expectation of missing data differs significantly from the true value.

2.2.1.6 K-Nearest Neighbours

The K-nearest neighbours (KNN) algorithm (Cover & Hart 1967) is a simple highly efficient algorithm that can be used to recover missing data (Kiani & Saleem 2017, Huang & Sun 2016, Pan & Li 2010). It replaces the missing value using the K-most similar non-missing values.

The KNN algorithm has three main steps. First, the distance between each complete sample and the sample containing missing values is calculated across the entire dataset. The Euclidean distance is often used to calculate this distance, the closer the samples are the higher the similarity between the samples. Second, the K samples closest to the samples containing missing values are selected. Third, the values are recovered by weighting and averaging these K samples.

The KNN algorithm can be adapted well for continuous data sets. Since the recovery process in the KNN algorithm is a self-learning process there is no need to learn a prediction model in advance. However, the accuracy of the recovery results depends on the artificially set parameter K. Some intelligent selection algorithms, such as grey

relation analysis (Huang & Sun 2016), have been proposed to address this issue. For large-scale datasets, the execution efficiency of the KNN algorithm is very low as it needs to traverse the entire dataset when performing data recovery on each missing datum. Thus, the main issue to address with the KNN algorithm is how to select the optimal hyperparameter K adaptively and improve model efficiency.

2.2.1.7 Multiple Imputation

A single imputed value cannot capture all the uncertainty about which value to implement. Multiple imputation (MI) (Rubin 1996, Meng & Rubin 1992, Rubin 2004, Schafer 1997) aims to address this by imputing multiple values rather than a single value, and then analysing each data set with standard methods. Multiple imputation is derived from Bayesian estimation, which considers that the values of missing data are random and originate from the known data. First, different methods are used to recover the missing data to form several complete datasets and reflect the uncertainty. Second, each complete dataset is analysed using a parameter estimation method. Finally, the missing values of the dataset are infilled by integrating multiple recovery results. Figure 2.2.1 shows the basic pipeline for MI method.

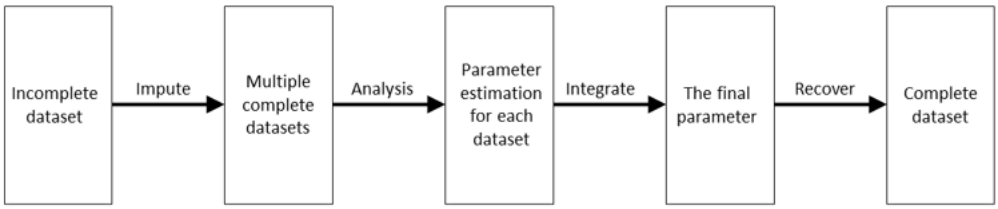


Figure 2.2.1: Basic pipeline for MI method (Du et al. 2020).

The MI method attempts to find potential relationships between missing variables and other variables by simulating the distribution of missing data as much as possible. The MI method maintains the uncertainty of the original dataset by using

the auxiliary information rationally. Compared to single-values imputation, multiple imputation can greatly improve recovery results (Faris et al. 2002).

2.2.1.8 Summary of Traditional Methods

Different data imputing methods are suitable for different cases. Table 2.2.1 summarises the traditional recovery methods describes and their suitability. Simple methods, such as the mean method or case deletion, are easy to implement, however, they are not best suited for large amounts of missing data. Methods such as EM or MI can result in high accuracy, however this is traded with high computational complexity. Many of the traditional methods are not suitable for large-scale data.

2.2.2 Non-Traditional Methods

While traditional methods are often limited for large-scale datasets, the development of methods for infilling missing data is an active field of research. Here we will give an overview of some of the common methods used in literature.

2.2.2.1 Support Vector Regression

Support vector regression (SVR) (Drucker· et al. 1996) is a regression algorithm that supports both linear and non-linear regressions. Compared to simple regression, where the aim is to minimise the error rate, SVR aims to find the error within a certain threshold to approximate the best value within a given margin. The advantages of SVR are it is relatively easy to implement, robust to outliers, reasonably computationally efficient and the prediction accuracy can be improved by measuring the confidence in the classification.

Data recovery methods based on SVR and extensions of SVR have been used in a variety of applications. Feng et al. (2005) implement a SVR approach for infilling missing SARS data. The SVR method was found to have high precision but required that the training set be complete. The size of the training set influenced the

Table 2.2.1: Summary of traditional methods for infilling missing data.

Methods	**Advantages**	**Disadvantages**	**Suitable Applications**
Deletion	Easy to implement.	Reduces sample size. Disrupts the structure of the data.	Low missing data ratio.
Mean Method	Straight forward calculation, the mean of the data stays the same.	Leads to bias of sample variance.	Large-scale datasets that approximately follow the normal distribution.
Nearest Neighbour linear interpolation	Straight forward calculation, easy to implement.	Leads to bias of sample variance.	Short periods of consecutive missing data.
Regression-based methods	Interpretable and gets good imputation accuracy due to effective modelling.	Requires prior knowledge	Only for univariate missing pattern, independent variables must be free of missing values.
Expectation maximisation (EM)	High imputation accuracy.	Convergence speed depends on size of samples.	Low missing data ratio.
K-Nearest Neighbours (KNN)	No model needed to learn in advance as self-learning.	Choice of K and distance metric affect accuracy. Efficiency depends on size of datasets.	Only for smaller datasets.
Multiple Imputation (MI)	Uncertainty of imputed value is considered in the modelling process.	Complex modelling process.	Only for smaller datasets.

accuracy of the predictions, as such as there needs to be a large enough complete training set available. Liu et al. (2015) proposed a SVR model combined with a genetic algorithm (GA) to address missing data due to sensor faults during waster gas monitoring. This missing data is estimated using a multiple input single output prediction model to include the multiple factors that influence waste gas concentration (e.g. spatial, temporal and environmental factors. Combing SVR with GA enhanced complementarity between sensors and improved the reliability of the monitoring system. However, the method had a long runtime and GA tends to find local optimums leading to early convergence. Further, Liu et al. (2016) introduced quantum genetic strategies and simulated annealing strategies into standard GA to solve the premature convergence and poor searchability in SVR. This method again enhanced complementarity between sensors and was more accurate than SVR, though it still struggled with long runtimes. Shang et al. (2018) consider a particle swarm optimisation based SVR for missing traffic data. This method had high input precision and good robustness. However, accuracy decreased with the amount of missing data and the method was sensitive to spatial temporal information.

2.2.2.2 Decision Trees

Decision trees (Twala 2009) are used for both regression and classification tests. Decision trees implement a hierarchical tree structure, which consists of a root node, branches, internal nodes and leaf nodes. Decision tree learning employs a divide and conquer strategy to identify optimal split points within a tree. Missing data recovery is carried out using this method by building decision trees to observe the missing values of each variable and, then fills the missing values of each variable by using its corresponding tree. The missing value prediction is shown in the leaf node. Decision trees can be used for both numerical and categorical data. Complex decision trees can be computationally expensive but often have a low bias.

There are several studies where decision trees are used to input missing data. Gimpy

et al. (2014) used a decision tree approach for estimating missing values from data mining. They demonstrated that the accuracy for the classification algorithm used on the data improved for the infilled data set compared to the incomplete dataset. However, no work on the size of the data set or mount of missing data was discussed. Rahman & Islam (2013) implement a decision tree and forest technique on nine different datasets for infilling missing values. Their results indicate their method has high accuracy but struggles with computational complexity and high memory usage. Rahman & Islam (2011) implement a regression tree approach alongside the EM-algorithm (see above). The EM algorithm improved the infilling on datasets with high correlation among attributes but in cases where the dataset was small it was often insufficient to get a good result from the EM algorithm.

2.2.2.3 Neural Network Based Methods

Neural networks (NN) are inspired by biological neural networks and can capture and make use of the temporal information. NNs are structured with an input layer that takes in the data and an output layer which generates the resulting outputs. Any layers in between do not see the input or output layer directly and are called hidden layers. Figure 2.2.2, shows the basic structure of a NN with one hidden layer. NN have been criticised for being non transparent due to the multilayer non-linear structure and outputs that are not traceable by humans (Buhrmester et al. 2021).

Since NN have the ability to build and learn models they can be applied to infilling missing data. (Coulibaly & Evora 2007) consider neural network methods for infilling missing daily weather records, comparing six NN. The methods considered were multilayer perception network, time lagged feedforward network, generalised radial basis function network, the recurrent neural network as well as the time delay neural network and the counter propagation fuzzy neural network. The multilayer perception network and time lagged feedforward network were the most accurate for both daily precipitation and extreme temperature records, whereas the generalised

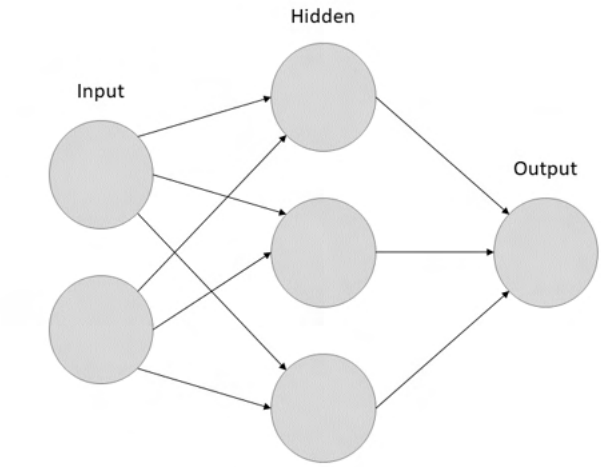


Figure 2.2.2: univariate missing

radial basis function network was most suitable for infilling minimum and maximum temperature but less suitable for precipitation values. This suggests the selection of NN is sensitive to the type of data used. Lee & Park (2015) consider an endpoint fixing method based feedforward neural network for missing data infilling for tide stations capturing water levels in coastal and ocean areas. This method was efficient and suitable in situations where nearly all tide stations simultaneous failed to record water levels However, the method was sensitive to missing window parameters.

Other types of NNs include self organizing maps (SOM), which are a tool for multivariate data, mapping data from a high dimensional input space to a low dimensional output space, serving as a clustering tool that can interpolate between previously encountered inputs. Experimental results show SOM can work well for both long and short gaps of missing data and results in higher accuracy compared to traditional methods and are less dependent on the location of the missing value (Junninen et al. 2004). The main limitation of SOM is that they have a long running time. Nkiaka et al. (2016) implement SOMs for infilling missing hydro-meteorological data, the methods were robust for infilling short gaps in hydro-meteorological time series. However, performance decreased for long periods of

consecutive missing data.

2.2.2.4 Ensemble Methods

Ensemble methods combine outputs from multiple models with the aim to produce a single improved result. Developing an ensemble involves creating varied models and merging their estimates, the aim being to have an ensemble that is as diverse as possible. Ensemble techniques are best suited to where the highest degree of accuracy is desired, studies have shown that ensemble based methods outperform single model approaches for imputing missing data (Zhang et al. 2019, Oehmcke et al. 2016). Ensemble methods can be parallelised which is practical for large data sets. Baruah et al. (2023) propose an ensemble technique for infilling missing data across a variety of datasets, ensembling k-nearest neighbour imputation, local least square imputation, miss forest imputation (a random forest imputation algorithm for missing data), and k-means clustering imputation. Their methods preformed best against compared methods the majority of the time, but not for all datasets. Their method was also noted to decrease in performance for high dimensional datasets. In general, ensemble methods tend to have long computational times as data needs to pass through multiple models and the strategy for building the best ensemble is dependent on the problem at hand Polikar (2006).

2.2.3 Summary of Non-traditional Methods

While non-traditional methods can produce excellent results for imputing missing data, they can be more challenging to implement, requiring a higher degree of skill to use. A common issue among the majority of the methods was long runtimes. This may result in methods being unsuitable for some applications regardless of accuracy if the computation time is too long. As well as the above methods, there are many more methods for infilling missing data than just those discussed here many methods, also many extensions or combinations of these methods all of which are used in various applications. Often these methods are compared against each

other and the 'best' method varies depending on the type of data, the missing data mechanism or pattern and how much data is missing. Emmanuel et al. (2021) compared multiple methods of missing data imputation including KNN, Random Forests and Ensemble methods concluding that the precision and accuracy depend strongly on the type of data being analysed and there is no clear indication that one method is preferred over another. Therefore, no one method is seen as the best tool for missing data infilling.

2.3 Conclusions

Missing data is an unavoidable aspect of monitoring systems, and the estimation of missing data is a critical task for improving the quality of the data and any analysis carried out on that data. Here, we considered the different types of missing data and several traditional methods and non-traditional methods for inputting missing data. While many tools exist for handling missing data, the performance of each method is dependent on the data size, the amount missing and the missing pattern. Additionally, the performance of each method depends on the application. While traditional methods are often easier to implement and have faster computation times, they tend to be less accurate than more sophisticated models. However, the more complex models trade higher accuracy for increased complexity and longer runtimes.

While a large amount of research has been carried out into infilling missing data, it remains an active field of study. For instance, a major investigation of this thesis is considering a time series missing data problem using surface level ozone measurements, meaning many of these methods are ill-suited as the data is not i.i.d (independently identically distributed). We aim to tackle the issue of long computation time by using a secondary dataset to infill the missing values. There exist many modelled environmental datasets that use physically-based models to

model the atmosphere at a given point. However, these datasets are often biased. We seek to implement a missing data method that can bias correct the modelled data and use this to estimate the missing values. By using a secondary dataset, we can keep the model relatively simple as the pollutant has already been modelled. This is carried out in Chapters 4 and 5, where we derive and implement a skew Kalman approach for infilling missing ozone data using bias corrected reanalysis data.

Chapter 3

Optimal Bayesian Filtering

In the previous Chapter, we reviewed methods for missing data. In order to set the foundations for the skew Kalman approach for missing data we will propose in Chapter 5, we must first introduce the traditional Kalman filter to provide the necessary background on which to derive the skew Kalman filter. Before deriving the skew Kalman filter, we will first consider the traditional Kalman filter. We observe the world through noisy measurements (e.g. sensors for air quality) and we want to be able to understand the underlying process of the system. Bayesian filters blend our noisy and limited knowledge of how a system behaves with limited sensor readings to produce an estimate of the state of the system.

3.1 Introduction State Space Models

A state space model is a discrete-time, stochastic model that contains two sets of equations, the state equation and the observation equation. Here we consider a hidden state $x_{0:T} = \{x_0, x_1, x_2, ..., x_T\}$ over time $t = \{0, 1, 2..., T\}$ that can only be observed through noisy measurements $\{y_1, y_2, ...\}$. This is shown in Figure 3.1.1. In optimal filtering, the goal is to estimate the hidden states $x_{0:T}$ from the observed measurements $y_{1:T}$. In the Bayesian sense, we want to compute the joint posterior distribution of all states given all the measurements.

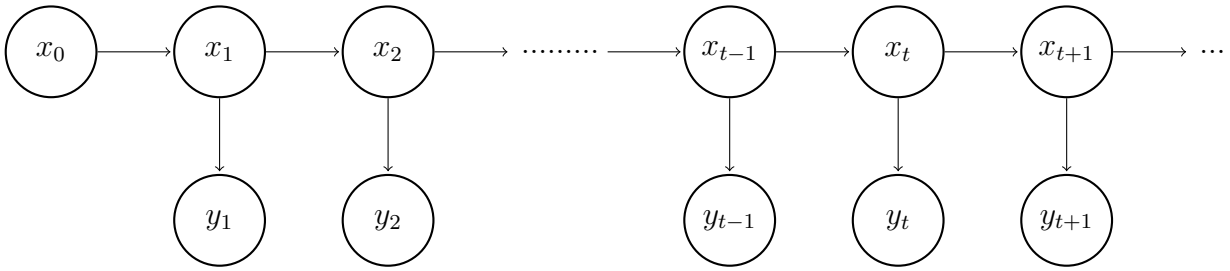


Figure 3.1.1: Dependence between variables in a state space model.

Definition 3.1.1. *(Probabilistic state space model) A probabilistic state space model, or non-linear filtering model, consists of a sequence of conditional probability distributions:*

$$\begin{aligned} x_t &\sim p(x_t|x_{t-1}), \\ y_t &\sim p(y_t|x_t), \end{aligned} \tag{3.1.1}$$

for t = 1,2,..., where

- $x_t \in \mathbb{R}^n$ *is the state of the system at time step t,*
- $y_t \in \mathbb{R}^m$ *is the measurement at time step t,*
- $p(x_t|x_{t-1})$ *is the dynamic model which describes the stochastic dynamics of the system.*
- $p(y_t|x_t)$ *is the measurement model, which is the distribution of measurements given the state.*

We assume the model to be Markovian and therefore the following two properties hold.

Property 3.1.1. *(Markov property of states) The states* $\{x_t : t = 0, 1, 2, ...\}$ *form a Markov sequence or Markov chain if discrete. The Markov property means that the evolution of* x_t *does not depend on past history before the time step t-1,*

$$p(x_t|x_{1:t-1}) = p(x_t|x_{t-1}), \tag{3.1.2}$$

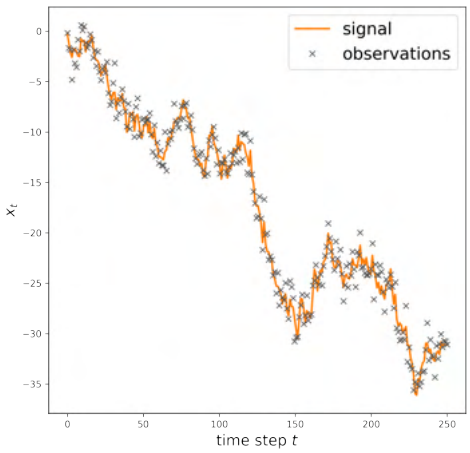


Figure 3.1.2: Gaussian random walk from Example 3.1.1 with parameters R=S=1.

Property 3.1.2. *(Conditional independence of measurements) The current measurement y_t given the current state x_t is conditionally independent of the measurement and state histories:*

$$p(y_t|x_{1:t}) = p(y_t|x_t) \tag{3.1.3}$$

Hence, we can think of x_t as a random process, which has simple Markovian dynamics, and we can assume that the observation y_t depends only on the system at the time of the measurement, x_t. Figure 3.1.1 illustrates the dependence of variables. The Gaussian random walk shown in Figure 3.1.2 is a Markovian sequence. We can combine this with measurements to obtain a state space model.

Example 3.1.1. (Gaussian random walk) We write the one-dimensional Gaussian random walk model as

$$\begin{aligned} x_t &= x_{t-1} + r_t, \qquad & r_t &\sim N(0, R), \\ y_t &= x_t + s_t, & s_t &\sim N(0, S). \end{aligned} \tag{3.1.4}$$

Using $R = S = 1$, Figure 3.1.2 shows an example of the signal x_t and the measurements y_t.

With the Markovian assumption and the filtering model (3.1.1), the joint prior distribution of the states $x_{0:T} = \{x_0, ..., x_T\}$ is given by,

$$p(x_{0:T}) = p(x_0) \prod_{t=1}^{T} p(x_t|x_{t-1}), \tag{3.1.5}$$

and the joint likelihood of the measurements $y_{1:T} = \{y_1, ..., y_T\}$ is,

$$p(y_{1:T}|x_{0:T}) = \prod_{t=1}^{T} p(y_t|x_t). \tag{3.1.6}$$

We would look to compute the posterior distribution of the hidden state conditional on the observations by Bayes' rule,

$$p(x_{0:T}|y_{1:T}) = \frac{p(y_{1:T}|x_{0:T})p(x_{0:T})}{p(y_{1:T})}, \tag{3.1.7}$$

where $p(y_{1:T})$ is the normalization constant defined as

$$p(y_{1:T}) = \int p(y_{1:T}|x_{0:T})p(x_{0:T})dx_{0:T}. \tag{3.1.8}$$

From Equation (3.1.7) we can see that each time a new measurement is added to the system the full posterior distribution would need to be recomputed. However, this is computationally expensive, alternatively, we can consider the marginal distribution of the states. The filtering distributions computed by the Bayesian filter are the marginal distributions of the current state x_t given the current and previous measurements $y_{1:T} = \{y_1, ..., y_T\}$:

$$p(x_t, y_{1:T}), \qquad t = 1, ..., T. \tag{3.1.9}$$

The prediction distributions are the marginal distributions of the future state x_{t+n}, n steps after the current time step:

$$p(x_{t+n}, y_{1:T}), \qquad t = 1, ..., T, \quad n = 1, 2, ... \tag{3.1.10}$$

3.2 Filtering Equations

We can write the state space model in the following form:

$$\begin{aligned} x_0 &\sim p(x_0), \\ x_t &\sim p(x_t|x_{t-1}), \\ y_t &\sim p(y_t|x_t), \end{aligned} \tag{3.2.1}$$

where we have

- an initial distribution which specifies the prior distribution $p(x_0)$ of the hidden state x_0 at the initial time step $t = 0$,
- a dynamic model which describes the system dynamics, and its uncertainties as a Markov sequence, defined in terms of the transition probability distribution $p(x_t|x_{t-1})$,
- a measurement model which describes how the measurements y_t depend on the current state x_t. This dependence is modelled by specifying the conditional probability distribution of the measurement given the state which is denoted as $p(y_t|x_t)$.

We denote the information provided by the first t observations by $D_t = \{y_1, ..., y_t\}$. The property of conditional independence of measurements and the Markov property of states can be used to compute the filtered and predictive densities using a recursive algorithm. Starting from $x_0 \sim p_0(x_0) = p(x_0|D_0)$ for $(t = 1, 2, ...)$ we can compute the following:

- the one-step ahead predictive density for x_t given D_{t-1}, based on the filtering density $p(x_{t-1}|D_{t-1})$ and the transition model $p(x_t|x_{t-1})$;
- the one-step ahead predictive density for the next observation $p(y_t|D_{t-1})$;
- the filtering density $p(x_t|D_t)$ using Bayes rule with $p(x_t|D_{t-1})$ and the likelihood of observation y_t, $p(y_t|x_t)$.

Proposition 3.2.1. *(filtering recursions).*

- *The predictive density for the states can be computed from the filtered density* $p(x_t|D_{t-1})$ *according to*

$$p(x_t|D_{t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|D_{t-1})dx_{t-1}. \tag{3.2.2}$$

- *The predictive density for the observations can be computed from the predictive density for the states as*

$$p(y_t|D_{t-1}) = \int p(y_t|x_t)p(x_t|D_{t-1})dx_t. \tag{3.2.3}$$

- *The posterior filtering density can be computed from the above densities as*

$$p(x_t|D_t) = \frac{p(y_t|x_t)p(x_t|D_{t-1})}{p(y_t|D_{t-1})}. \tag{3.2.4}$$

Proof. Looking first at the predictive density for the states,

$$\begin{aligned} p(x_t|D_{t-1}) &= \int p(x_{t-1}, x_t|D_{t-1})dx_{t-1}, \\ &= \int p(x_t|x_{t-1}, D_{t-1})p(x_{t-1}|D_{t-1})dx_{t-1}, \\ &= \int p(x_t|x_{t-1})p(x_{t-1}|D_{t-1})dx_{t-1}, \end{aligned} \tag{3.2.5}$$

from the conditional independence $x_t \perp\!\!\!\perp (y_1, ..., y_{t-1})|x_{t-1}$.

Again, making use of the assumption of conditional independence, for the predictive density for the observations we can use the conditional independence $y_t \perp\!\!\!\perp (y_1, ..., y_{t-1})|x_t$ to get

$$\begin{aligned} p(y_t|D_{t-1}) &= \int p(y_t, x_t|D_{t-1})dx_t, \\ &= \int p(y_t|x_t, D_{t-1})p(x_t|D_{t-1})dx_t, \\ &= \int p(y_t|x_t)p(x_t|D_{t-1})dx_t. \end{aligned} \tag{3.2.6}$$

The filtering density can be found using Bayes rule and the conditional independence $y_t \perp\!\!\!\perp (y_1, ..., y_{t-1})|x_t$

$$\begin{aligned} p(x_t|D_t) &= \frac{p(x_t, D_t)}{p(D_t)} = \frac{p(x_t|y_t, D_{t-1})}{p(y_t|D_{t-1})p(D_{t-1})} = \frac{p(x_t|D_{t-1})p(y_t|x_t, D_{t-1})}{p(y_t|D_{t-1})} \\ &= \frac{p(y_t|x_t)p(x_t|D_{t-1})}{p(y_t|D_{t-1})} \end{aligned} \tag{3.2.7}$$

□

We can then compute recursively the k-steps ahead predictive densities, starting for $k = 1$ using

$$\begin{aligned} p(x_{t+k}|D_t) &= \int p(x_{t+k}|x_{t+k-1})p(x_{t+k-1}|D_t)dx_{t+k-1}, \\ p(y_{t+k}|D_t) &= \int p(y_{t+k}|x_{t+k})p(x_{t+k}|D_t)dx_{t+k}, \end{aligned} \tag{3.2.8}$$

where $p(x_{t+k}|D_t)$ summarises the information contained in the past observation D_t, this is sufficient for predicting y_{t+k}. In general, the filtering equations do not have a closed-form expression, except for the Kalman filter, where the state and observation processes are linear-Gaussian. We will consider the Kalman filter in the following section.

3.3 Kalman Filter

There are many examples in which data are noisy or partially recorded, where we want to understand the true underlying process of the data, e.g. global positioning system (GPS), target tracking, and multiple target tracking (Särkkä 2013). We can solve this problem using a state-space model framework, where we assume there exists a latent process that evolves linearly with Gaussian noise, and which is observed indirectly with additional Gaussian noise. This type of state space model is known as the Kalman Filter (Kalman 1960). Kalman filtering is an efficient recursive algorithm for tracking a time-dependent state vector in real-time with a noisy evolution equation and noisy measurements. Specifically, the Kalman filter is the closed-form solution of the Bayesian filtering equations for the filtering model,

$$x_t = A_t x_{t-1} + r_t, \tag{3.3.1}$$

$$y_t = H_t x_t + s_t, \tag{3.3.2}$$

where $x_t \in \mathbb{R}^n$ is the state, $y_t \in \mathbb{R}^m$ is the measurement, $r_t \sim N(0, R_t)$ is the process noise, and $s_t \sim N(0, S_t)$ is the measurement noise. The prior distribution, $x_0 \sim N(m_0, P_0)$, is Gaussian, the matrix A_t represents the transition matrix of the dynamic linear model, and H_t is the measurement model matrix. The first equation is called the state equation and the second is called the observation equation. We can rewrite Equations (3.3.1) and (3.3.2) in terms of their distributional form as:

$$\begin{aligned} p(x_t|x_{t-1}) &= N(x_t|A_t x_{t-1}, R_t), \\ p(y_t|x_t) &= N(y_t|H_t x_t, S_t). \end{aligned} \tag{3.3.3}$$

It follows from the properties of Gaussian distributions that the marginal and conditional distributions are also Gaussian. We seek to learn the distribution of the latent process x_t, conditional on the noisy observations recorded so far, D_t. The distribution $p(x_t|D_t)$ is referred to as the filtering distribution, and likewise,

$p(x_t|D_{t-1})$ and $p(y_t|D_{t-1})$ are the predictive distribution and marginal likelihood, respectively. The solution to the filtering problem is given by the Kalman filter.

Theorem 3.3.1. *(Kalman filter) The resulting distributions from evaluating the Bayesian filtering equations for the linear filtering model are Gaussian. For the dynamic linear model Equations* (3.3.1) *and* (3.3.2)*, if*

$$x_{t-1}|D_{t-1} \sim N(m_{t-1}, P_{t-1}), \tag{3.3.4}$$

where $t \geq 1$*, then*

- *the one-step ahead predictive density of* x_t*, given* D_{t-1} *Equation* (3.2.2)*, is Gaussian with parameters*

$$\begin{aligned} m_{t|t-1} &= E(x_t|D_{t-1}) = A_t m_{t-|t-1}, \\ P_{t|t-1} &= Var(x_t|D_{t-1}) = A_t P_{t-1|t-1} A_t' + R_t, \end{aligned} \tag{3.3.5}$$

- *the one-step ahead predictive density of* y_t *given* D_{t-1} *Equation* (3.2.3)*, is Gaussian, with parameters*

$$\begin{aligned} f_t &= E(y_t|D_{t-1}) = H_t m_{t|t-1}, \\ Q_t &= Var(y_t|D_{t-1}) = H_t P_{t|t-1} H_t' + S_t, \end{aligned} \tag{3.3.6}$$

- *the filtering density of* x_t *given* D_t *Equation* (3.2.4)*, is Gaussian with*

$$\begin{aligned} m_{t|t} &= E(x_t|D_t) = m_{t|t-1} + P_{t|t-1} H_t' Q_t^{-1} e_t, \\ P_t &= Var(x_t|D_t) = P_{t|t-1} - P_{t|t-1} H_t' Q_t^{-1} H_t P_{t|t-1}, \end{aligned} \tag{3.3.7}$$

where $e_t = (y_t - f_t)$ *is the forecast error.*

Proof. The joint density for a random variable $(x_0, x_1, ..., x_n, y_1, ..., y_n)$ for any $n \geq 1$

is given by,

$$(x_0, x_1, ..., x_n, y_1, ..., y_n) \sim p_0(x_0) \prod_{t=1}^{n} p(y_t|x_t)p(x_t|x_{t-1}). \tag{3.3.8}$$

As the process $((x_t, y_t), t = 1, 2, ...)$ is Markovian, we can make use of Properties 3.1.1 and 3.1.2 in Equation (3.3.8). In the linear Gaussian setting, the marginal and conditional distributions are Gaussian. The Kalman filter can be derived using some of the properties of Gaussian distributions given in Lemma A.1.1 and A.1.2, hence the joint density of $(x_0, x_1, ..., x_n, y_1, ..., y_n)$ for any $t \geq 1$ is Gaussian, as is the conditional distribution of some component given another component. Then it follows that the predictive densities and the filtering densities, Proposition 3.2.1, are Gaussian and we can compute their means and variances. If $x_{t-1} \sim N(m_{t-1,P_{t-1}})$ then,

- from the state space equation, $x_t|D_{t-1} \sim N(m_{t|t-1}, P_{t|t-1})$, where

$$\begin{aligned} m_{t|t-1} &= E(x_t|D_{t-1}) = E(E(x_t|x_{t-1}, D_{t-1})|D_{t-1}) = E(A_t x_{t-1}|D_t), \\ &= A_t m_{t-1|t-1}. \\ P_{t|t-1} &= Var(x_t|D_{t-1}) = E(Var(x_t|x_{t-1}, D_{t-1})|D_{t-1}) \\ &+ Var(E(x_t|x_{t-1}, D_{t-1})|D_{t-1}), \\ &= R_t + A_t P_{t-1|t-1} A_t'. \end{aligned} \tag{3.3.9}$$

- From the observation equation, $y_t|D_{t-1} \sim N(f_t, Q_t)$

$$\begin{aligned} f_t &= E(y_t|D_{t-1}) = E(E(y_t|x_{t-1}, D_{t-1})|D_{t-1}) \\ &= E(F_t x_{t-1}|D_t), \\ &= F_t m_{t|t-1}. \\ Q_t &= Var(y_t|D_{t-1}) = E(Var(y_t|x_{t-1}, D_{t-1})|D_{t-1}) \\ &+ Var(E(y_t|x_{t-1}, D_{t-1})|D_{t-1)}), \\ &= S_t + H_t P_{t|t-1} H_t'. \end{aligned} \tag{3.3.10}$$

- Using Bayes rule for computing the conditional density of $x_t|D_t$, with the density $N(m_{t|t-1}, P_{t|t-1})$ of $x_t|D_{t-1}$ as the prior and the density $N(H_t x_t, V_t)$ of $y_t|x_t$ as the likelihood,

$$y_t = H_t x_t + s_t, \qquad v_t \sim N(0, S_t), \tag{3.3.11}$$

where the parameters of x_t have a conjugate Gaussian prior $N(m_{t|t-1}, P_{t|t-1})$ and S_t are known. Then we have that

$$x_t|D_t \sim N(m_{t|t}, P_{t|t}), \tag{3.3.12}$$

where

$$\begin{aligned} m_{t|t} &= m_{t|t-1} + P_{t|t-1} H_t' Q_t^{-1} (y_t - H_t m_{t|t-1}), \\ P_{t|t} &= P_{t|t-1} + P_{t|t-1} H_t' Q_t^{-1} H_t P_{t|t-1}. \end{aligned} \tag{3.3.13}$$

□

The recursion is started from the prior mean m_0 and covariance P_0, then computing $p(x_1|D_1)$ and proceeding recursively as new data becomes available. The conditional density of $x_t|D_t$ solves the filtering problem. Typically, we are interested in a point estimate which is given by the conditional expected value $m_t = E(x_t|D_t)$. From the Kalman filter, m_t can be expressed in terms of the prediction mean $m_{t|t-1}$ plus a correction term. The correction term is given by the Kalman gain matrix, K, which gives a weighting between the new observation and the estimate:

$$\begin{aligned} K_t &= P_{t|t-1} H_t' Q_t^{-1}, \\ &= P_{t|t-1} H_t' (S_t + H_t P_{t|t-1} H_t')^{-1}. \end{aligned} \tag{3.3.14}$$

The weight of current information y_t depends on the observations covariance matrix S_t and on the predicted covariance matrix $P_{t|t-1} = Var(x_t|D_{t-1}) = R_t + A_t P_{t-1|t-1} A_t'$. Using the Kalman gain matrix, often the Kalman filter is summarised as two steps:

the time update and measurement update. Starting from the prior mean m_0 and covariance P_0, the equations from these steps are outlined below.

- Time update equations. The prior state and error covariance estimates are obtained by

$$\begin{aligned} m_{t|t-1} &= A_t m_{t-1|t-1}, \\ P_{t|t-1} &= A_t P_{t-1|t-1} A' + R_t. \end{aligned} \tag{3.3.15}$$

- Measurement update equations. The Kalman gain, posteriori state, and posteriori error covariances are obtained by

$$\begin{aligned} K_t &= P_{t|t-1} H_t' [H_t P_{t|t-1} H_t' + S_t]^{-1}, \\ m_{t|t} &= m_{t|t-1} + K_t (y_t - H_t m_{t|t-1}), \\ P_{t|t} &= (I - K_t H_t) P_{t|t-1}. \end{aligned} \tag{3.3.16}$$

Example 3.3.1. (Kalman filter for a Gaussian random walk) Applying the Kalman filter to the observations y_t given by the Gaussian random walk in Example 3.1.1 to estimate the state x_t at each time step gives the following

- Time update step

$$\begin{aligned} m_{t|t-1} &= m_{t-1|t-1}, \\ P_{t|t-1} &= P_{t-1|t-1} + R_t. \end{aligned} \tag{3.3.17}$$

- Measurement update step

$$\begin{aligned} K_t &= (P_{t|t-1})(P_{t|t-1} + S_t)^{-1}, \\ m_{t|t} &= m_{t|t-1} + K_t (y_t - m_{t|t-1}), \\ P_{t|t} &= P_{t|t-1} - K_t P_{t|t-1}. \end{aligned} \tag{3.3.18}$$

Figure 3.3.1 shows the Kalman filter applied to the Gaussian random walk. The filter is showing the underlying process of the data reasonably well, with a mean squared error of 0.6 when comparing the underlying process and Kalman filter estimate.

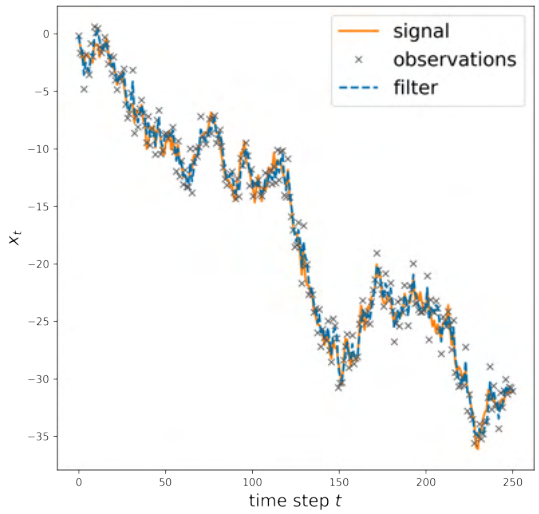


Figure 3.3.1: Kalman filter applied to the Gaussian random walk from Example 3.1.1 with parameters R=S=1.

We can also use this example to better understand the role of the Kalman gain matrix and how it applies a weighting between the measurement and estimated value. The signal-to-noise ratio, given by the ratio between the two error variances $r = R/S$, greatly influences the behaviour of the process y_t. We can express $m_{t|t} = K_t y_t + (1 - K_t) m_{t-1|t-1}$ as a weighted average of y_t and m_{t-1} with weight

$$K_t = \frac{P_{t|t-1}}{Q_t} = \frac{P_{t-1|t-1} + R}{(P_{t-1|t-1} + R) + S}, \qquad 0 < K_t < 1. \tag{3.3.19}$$

If the signal-to-noise ratio, $r = R/S$ is small, then K_t is also small since S is larger than R meaning the measurement noise is larger and y_t is given little weight. If $S = 0$, we have the one-step ahead forecast given by the most recent data point as all the weighting goes to y_t since $K_t = 1$.

The Kalman filter steps are shown in Figure 3.3.2. The filter is started with an initial state m_0, and covariance matrix P_0 which feeds in from the previous step, from this a new state $m_{t|t-1}$ and covariance matrix $P_{t|t-1}$ are calculated. Next, the Kalman gain matrix K_t is calculated and the new predicted state is updated with the measurement at the current time step and the Kalman gain is used to give the new state estimate $m_{t|t}$. The Kalman gain is then used to find the new covariance matrix $P_{t|t}$ at the current time step. The new state and covariance matrix are then entered as the previous state for the process to repeat.

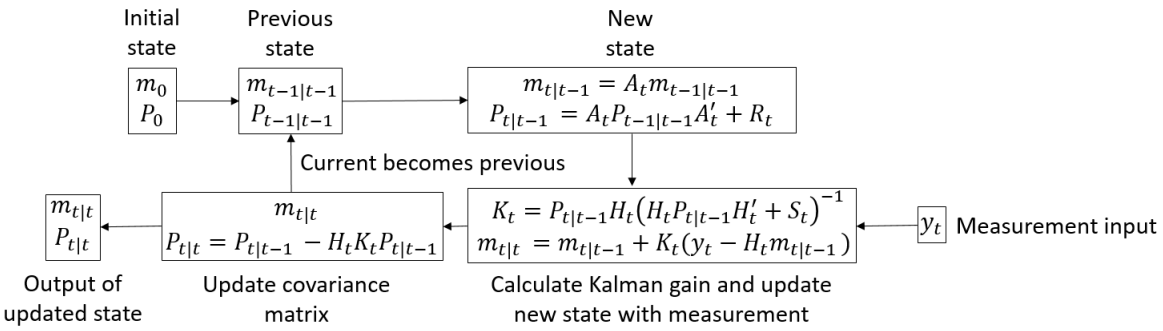


Figure 3.3.2: Recursive Kalman filter algorithm

3.4 Parameter Estimation in State Space Models

In state space models there are often unknown parameters that should be estimated along with the state itself.

3.4.1 Bayesian Approach

A state space model with unknown parameters $\theta \in \mathbb{R}^d$, modelled as random variables with a certain prior distribution $p(\theta)$, can be written in the form (Särkkä 2013)

$$\begin{aligned} \theta &\sim p(\theta), \\ x_0 &\sim p(x_0|\theta), \\ x_t &\sim p(x_t|x_{t-1}, \theta), \\ y_t &\sim p(y_t|x_t, \theta), \end{aligned} \tag{3.4.1}$$

using Bayes' rule the joint posterior distribution is

$$p(x_{0:T}, \theta|y_{1:T}) = \frac{p(y_{1:T}|x_{0:T}, \theta)p(x_{0:T}|\theta)p(\theta)}{p(y_{1:T}|\theta)}, \tag{3.4.2}$$

where

$$p(x_{0:T}, \theta) = p(x_0|\theta)\prod_{t=1}^{T} p(x_t|x_{t-1}, \theta), \tag{3.4.3}$$

and from the Markov property of states,

$$p(y_{1:T}|x_{0:T}, \theta) = \prod_{t=1}^{T} p(y_t|x_t, \theta). \tag{3.4.4}$$

If we are only interested in the parameters θ, then the marginal posterior of parameters given by integrating out the states is

$$p(\theta|y_{1:T}) = \int p(x_{0:T}, \theta|y_{1:T})dx_{0:T}. \tag{3.4.5}$$

However, this integral would be computationally challenging to compute, increasing in complexity as we obtain more measurements. Alternatively, a recursive maximum likelihood approach is faster and discussed in the next section.

3.4.2 Maximum Likelihood Estimation

The parameters in the state space model Equations (3.3.1) and (3.3.2) can be estimated using θ to represent the vector of unknown parameters in the initial mean m_0, initial covariance P_0, the transition matrix A, and the state and observation covariance matrices R and S. The likelihood function is given by the joint density of the observation for a particular value of the parameter, $p(y_1, ..., y_n; \theta)$, up to a constant factor such that $L(\theta) = c \times p(y_1, ..., y_n; \theta)$, where c is a constant. For the dynamic linear model, it is convenient to write the joint density of the observations in the form (Petris et al. 2009)

$$p(y_1, ..., y_n; \theta) = \prod_{t=1}^{n} p(y_t | D_{t-1}; \theta), \tag{3.4.6}$$

where $p(y_t|D_{t-1}; \theta)$ is the conditional density of y_t given the data up to time $t-1$, assuming that θ is the value of the unknown parameter. Since we know the terms in the RHS of Equation (3.4.6) are Gaussian densities with mean f_t and variance Q_t, Theorem 3.3.1, we can write the log-likelihood as

$$l(\theta) = -\frac{1}{2}\sum_{t=1}^{n} \log |Q_t(\theta)| - \frac{1}{2}\sum_{t=1}^{n} (y_t - f_t(\theta))' Q_t(\theta)^{-1} (y_t - f_t(\theta)), \tag{3.4.7}$$

which can be numerically maximised to obtain the maximum likelihood estimator (MLE) of θ. Computationally, rather than maximising the likelihood function it is equivalent to minimising the negative of the likelihood. The log-likelihood function can be minimised by setting up a set of recursions for the log-likelihood function and its first two derivatives. Newton's method can be used to update the parameters until the negative log-likelihood is minimised. Newton's method looks to find the root of a continuous, differentiable function $l(\theta)$ by looking at a point close to the root and obtaining a better estimate from it. Consider the point $\theta = \theta^{(0)}$ which you know to be near the root, then using Newton's method a better estimate $\theta^{(1)}$ is

given by

$$\theta^{(1)} = \theta^{(0)} - \frac{l(\theta^{(0)})}{l'(\theta^{(0)})}. \tag{3.4.8}$$

This can be carried out iteratively,

$$\theta^{(k)} = \theta^{(k-1)} - \frac{l(\theta^{(k-1)})}{l'(\theta^{(k-1)})}, \tag{3.4.9}$$

until a desired level of accuracy is met.

The parameters in the Kalman filter can be estimated with the following steps using Newton's method:

1. Set initial parameters, $\theta^{(0)}$.

2. Run the Kalman filter with initial parameters, to obtain the forecast errors $y_t - f_t(\theta)$ and error covariances Q_t.

3. Minimise the negative log-likelihood Equation (3.4.7), with Newton's method, to obtain estimates for the parameters $\theta^{(1)}$.

4. Repeat steps 2 and 3 with the new parameters until the estimates or likelihood stabilise.

Example 3.4.1. Consider a similar case to Example 3.1.1. We have

$$\begin{aligned} y_t &= x_t + s_t, & s_t &\sim N(0, S) \\ x_t &= x_{t-1} + r_t, & r_t &\sim N(0, R) \end{aligned} \tag{3.4.10}$$

where s_t, r_t, and x_0 are independent, and $t = 1, 2, ...$ Using the same notation we have, $R = 0.5$ and $S = 1$. The parameter estimation was accomplished using the *optimize* function in Python. The final estimates for R=0.46 and $S = 1.04$ in 31 iterations.

3.5 Kalman Filtering for Bias Correction and Recovering Missing Data

Kalman filters and extensions of are used in a variety of applications, including i) post-processing, where Kalman filters are used after forecasts have been made using climate models to reduce the errors and improve the accuracy of the results (Djalalova et al. 2015, Heemink & Segers 2002, Ridder et al. 2012); ii) forecasting using observational data and suitable covariates (Hoi et al. 2008, Zolghadri & Cazaurang 2006); and iii) inverse modelling, where the model parameters used to produce the data are determined (Napelenok et al. 2008, Metia et al. 2020).

There are many examples of the Kalman filter in various forms that are used for bias correction and recovering missing data. The Kalman filter has been used to improve the analysis and prediction results of ground ozone concentrations by reconstructing the emissions. The results of the analysis procedure were found to be less sensitive to fluctuations in the data (Heemink & Segers 2002). Kalman filtering has also been used along with historical forecast analogues to improve surface PM forecasts from air quality models that contained large, seasonally varying biases (Djalalova et al. 2015). Kalman filter approaches have been used alongside air quality models, resulting in a reduction in the error of ozone estimation at measurement sites (Agudelo et al. 2011). The Kalman filter is also used to reduce the bias and lower the error between estimation and measurements by applying a Kalman filter-based bias correction method to the output from air quality models (Ridder et al. 2012). Inverse modelling methods using Kalman filters are also used to identify biases in emission inventories using satellite observations (Napelenok et al. 2008).

The next Chapter will consider an extension to the Kalman filter that allows for skewness to be present in the observation noise. This additional information incorporated from the skewness can be used to capture the bias between two

datasets to provide a Kalman filtering approach for bias correction. In Chapter 5 we implement the skew Kalman filter using ozone measurement data and reanalysis data to both correct the bias between two datasets and recover missing data on the unbiased dataset.

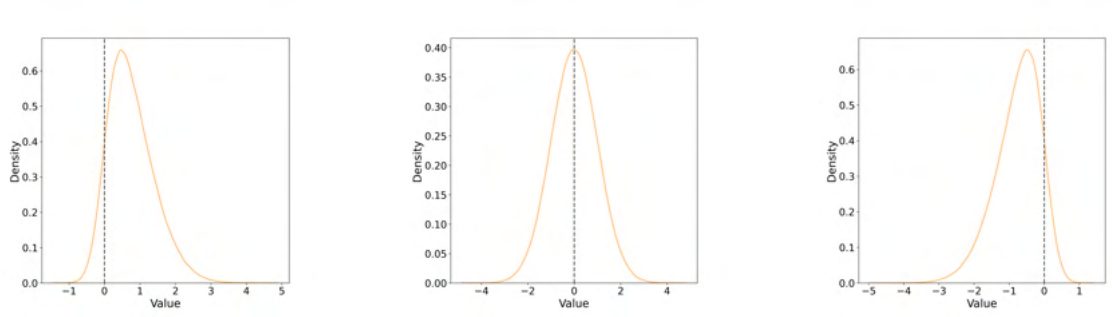
Chapter 4

The Skew Kalman Filter

Following from the previous chapter which discussed the background for the Kalman filter. We will now look at extending the Kalman filter to allow for skewness to be present in the observation noise. This allows for additional flexibility compared to the traditional filter.

4.1 Introduction

As discussed in the previous chapter, the Kalman filter is a useful tool for analysing and forecasting time series data. One key assumption of the Kalman filter is normality in the observation noise. This allows for the Kalman filtering steps to be performed quickly and efficiently, since the multivariate normal distribution is completely characterised by its mean and variance. When this normality assumption holds, the properties of the multivariate normal means the Kalman filter is tractable. When the observation noise is normally distributed the measurements are assumed to be symmetric about the mean. However, if there are biases present, this assumption may not hold, for example, if the measurements of the system are consistently too high or too low compared to the true value, the observations will not be symmetric about the mean. This non symmetric distribution about the mean can be captured using a skew distribution, this is shown in Figure 4.1.1. We seek to capture this bias



(a) Skew normal distribution with mean = 0, variance = 1 and skew = 3.

(b) Skew normal distribution with mean = 0, variance = 1 and skew = 0.

(c) Skew normal distribution with mean = 0, variance = 1 and skew = -3.

Figure 4.1.1: Distribution plots of **(a)** a skew normal distribution with positive skew, **(b)** a normal distribution (i.e. no skew present), and **(c)** a skew normal distribution with negative skew. The mean is shown as a dashed line in each plot.

as a skewness on the observation noise such that we can use a skew Kalman filter to correct for that skew. This chapter proposes a simplified version of the skew normal Kalman filter that follows a multivariate skew normal distribution.

The skew normal Kalman filter is an extension to the typical Kalman filter (see Chapter 3) that accounts for the skewness present in the observational data to produce more accurate forecasts. While the skew Kalman filter was first proposed in 2005 (Naveau et al. 2005), it has failed to gain the same popularity in applications. This is despite its additional usefulness in that it removes the limit that the data need to be normally distributed and so reduces the need to transform data. Incorporating skewness into the traditional Kalman filter aims to improve its applicability by extending its usefulness to a wider range of data distributions but without compromising its low cost computational benefits. The benefits of a skew normal Kalman filter have been demonstrated using simulated data, as well as with a limited number of applications (Arellano-Valle et al. 2019, Fasano et al. 2019).

Current versions of the skew Kalman filter use an extension to the multivariate skew normal, either the univariate skew normal (SUN) (Arellano-Valle & Azzalini 2006)

or closed skew normal distributions (González-Farías et al. 2004). Compared to the multivariate skew normal family, which preserves just the marginal distribution, the univariate skew normal also preserves the conditional distributions (Arellano-Valle & Azzalini 2020). Estimating the parameters can be more computationally challenging than the basic filter due to the use of the unified skew normal distribution. Here, we develop a simplified approach to the skew Kalman filter, building on that proposed by Arellano-Valle et al. (2019), that reduces the unified skew normal to a standard skew normal. This results in a skew filter that is more similar to the basic filter, such that we are simply left with basic filtering equations plus an additional term. The main advantage of using our simplified approach is seen in the estimation of the model parameters, which is typically more complex and computationally expensive in the skew Kalman filter.

This chapter is outlined as follows, Section 4.2 introduces the skew Kalman filter, the challenges with estimating the parameters using maximum likelihood estimation and presents our simplified approach to the skew Kalman filter to address this issue. Section 4.3 is a simulation study, looking at varying levels of skewness in the observation noise and evaluates the skew Kalman filters performance compared to the standard Kalman filter. Finally, we give our conclusions in Section 4.4.

4.2 Skew Normal Kalman Filter

The traditional Kalman filter performs well under Gaussian noise. However, the observation noise is not always Gaussian.

Example 4.2.1. (Gaussian random walk with skewed noise) Looking again at the Gaussian random walk from Example 3.1.1, we can now include skewness in the

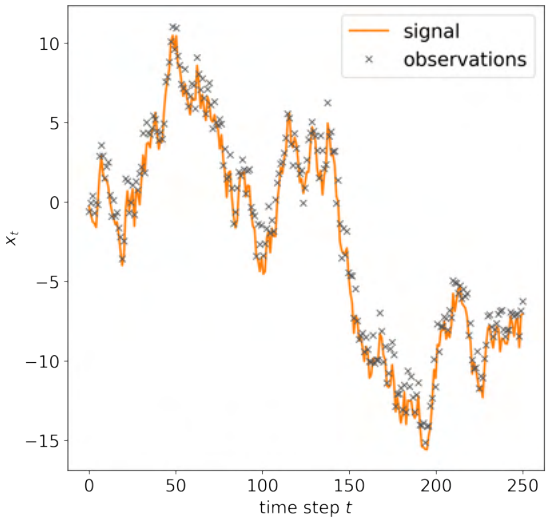


Figure 4.2.1: Gaussian random walk from Example 4.2.1 with parameters $W = V = 1$ and $\lambda = 2$.

observation noise and write the one-dimensional Gaussian random walk model as

$$\begin{aligned} x_t &= x_{t-1} + w_t, \qquad & w_t &\sim N(0, W), \\ y_t &= x_t + v_t, & v_t &\sim SN(0, V, \lambda), \end{aligned} \tag{4.2.1}$$

where $SN(0, V, \lambda)$ denotes a skew normal with location parameter 0, scale V and skew λ. Using $W = V = 1$ and $\lambda = 2$, Figure 4.2.1 shows an example of the signal x_t and the measurements y_t. Figure 4.2.2 shows the Kalman filter applied to the first 100 time steps of the data shown in Figure 4.2.1. Applying the traditional Kalman filter to these data results in the filter estimate consistently being higher than the signal. This is the result of the filter trying to fit to the mean of the data. However, due to the skewness present in the data the underlying process is lower than the mean. Therefore, we need to be able to account for this skewness in our Kalman filter to be able to capture the underlying signal.

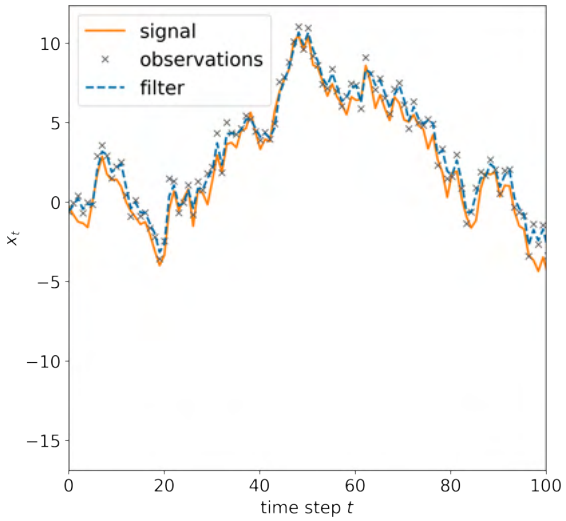


Figure 4.2.2: The standard Kalman filter applied to the Gaussian random walk from Example 4.2.1 with parameters $W = V = 1$ and $\lambda = 2$.

4.2.1 The Skew Normal Distribution

Before deriving the simplified skew Kalman filter, we will explore the issues currently present with the skew Kalman filter in the literature when using maximum likelihood estimation to estimate the parameters.

The unified skew normal (SUN) or closed skew normal distributions have been the preferred choice for the skew Kalman filter as compared to the multivariate skew normal family which preserves just the marginal distribution, the unified skew normal also preserves the conditional distributions (Arellano-Valle & Azzalini 2020). However, estimating the parameters can be more computationally challenging than the basic filter due to the use of the unified skew normal distribution. Following Arellano-Valle & Azzalini (2006), a random vector $X \in \mathbb{R}^p$ has a multivariate SUN distribution,

$$X \sim SUN_{p,q}(\xi, \Omega, \Lambda, \tau, \Gamma), \tag{4.2.2}$$

with location vector $\xi \in \mathbb{R}$, positive definite scale matrix $\Omega \in \mathbb{R}^{p \times p}$, skewness/shape matrix $\Lambda \in \mathbb{R}^{q \times p}$, extension vector $\tau \in \mathbb{R}^q$ and positive definite extension matrix

$\Gamma \in \mathbb{R}^{q \times q}$. The pdf is given by

$$f(x|\xi, \Omega, \Lambda, \tau, \Gamma) = \frac{\phi_p(x|\xi, \Omega)\Phi_q(\Lambda(x-\xi)+\tau|\Gamma)}{\Phi_q(\tau|\Gamma+\Lambda\Omega\Lambda')}, \qquad x \in \mathbb{R}^p. \tag{4.2.3}$$

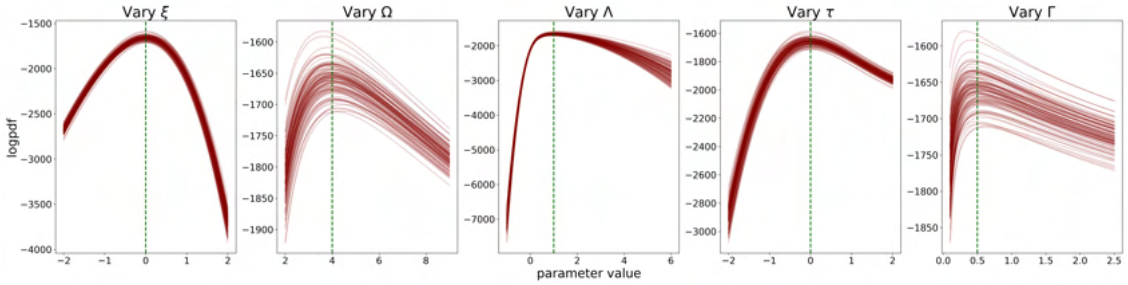
Where, $\phi(\cdot)$ and $\Phi(\cdot)$ represent a normal probability density function (pdf) and cumulative density function (cdf), respectively. The unified skew normal is difficult to estimate using maximum likelihood estimation. Sampling from Equation (4.2.2), the distribution parameters can only be estimated for certain parameter values. Figure 4.2.3 shows log likelihood plots for the parameters when $\tau = 0$, $\tau = 1$, and $\tau = -1$. The remaining parameters are kept fixed for each value for τ. While the parameters can be correctly estimated when $\tau = 0$, this does not hold well for other values of τ. If τ is limited to 0, it is logical to simplify to the multivariate skew normal, as when $q = 1$, $\tau = 0$ and $\Lambda = \eta$ Equation (4.2.3) simplifies to the multivariate SN pdf,

$$f(x|\xi, \Omega, \eta) = 2\phi_p(x|\xi, \Omega)\Phi_1(\eta(x-\xi)|\Omega) \tag{4.2.4}$$

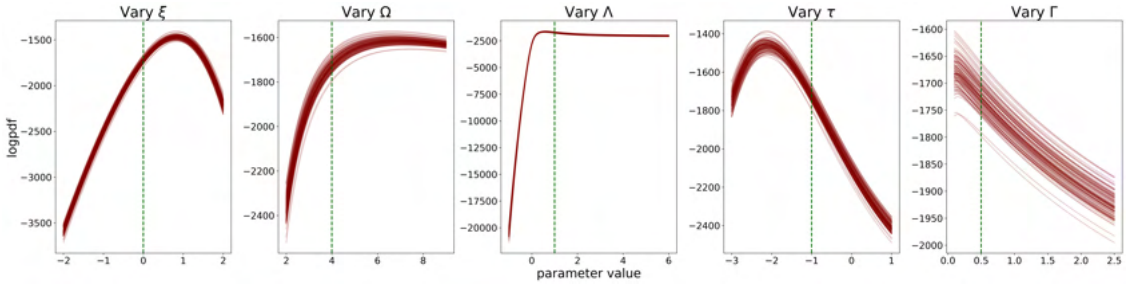
Thus, we will present a simplified approach to the unified skew normal Kalman filter, as proposed by Arellano-Valle et al. (2019). We can simplify the unified skew normal to a multivariate skew normal by fixing $q = 1$, $\tau = 0$ and $\Lambda = \eta$, allowing the filtering equations to also follow a multivariate skew normal and thus drastically reducing the computational complexity of the pdf. The computational complexity of the SUN pdf is due to the two multivariate normal cdf terms in the pdf. As a result, the filtering equations for the unified skew normal Λ, τ and Γ increase in dimension with time and thus the cdf terms quickly become computationally intractable.

4.2.2 Filtering Equations

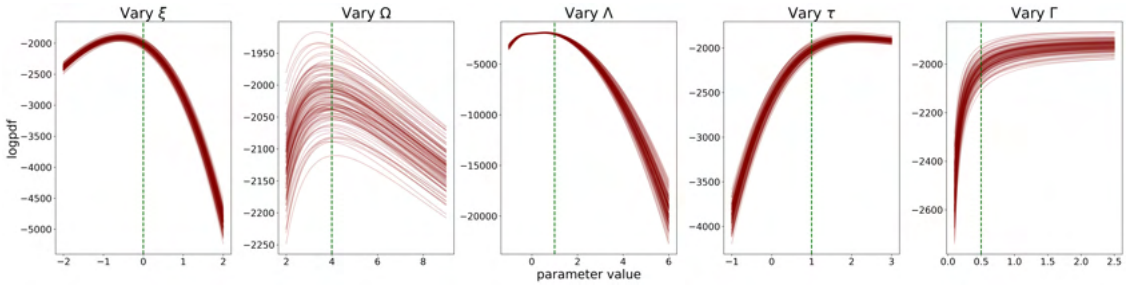
Following Arellano-Valle et al. (2019), the linear model given by Equations (3.3.1) and (3.3.2), can be represented using the skew-normal dynamic linear model



(a)



(b)



(c)

Figure 4.2.3: Log likelihood plots for 100 data sets sampled from Equation (4.2.2) with $\xi = 0$, $\Omega = 4$, $\lambda = 1$, $\Gamma = 0.5$ and **(a)** $\tau = 0$, **(b)** $\tau = -1$, and **(c)** $\tau = 1$. The dashed green line indicated the true values of each parameter. Each plot shows on parameter varied while the remaining parameters are fixed at the correct value.

(SNDLM), with state θ_t and measurement y_t, is given by

$$\theta_t|\theta_{t-1}, W_t \sim N_p(G_t\theta_{t-1}, W_t), \tag{4.2.5}$$

$$y_t|\theta_t, V_t, \lambda_t \sim SN_n(F_t'\theta_t - \sqrt{\frac{2}{\pi}}\Delta_t, V_t, \lambda_t), \tag{4.2.6}$$

where $\Delta_t = V_t\eta_t/(1+\eta_t'V_t\eta_t)^{1/2} = V_t^{1/2}\delta_t$, $\eta_t = V_t^{-1/2}\lambda_t$, $\delta_t = \lambda_t/(1+\lambda_t'\lambda_t)^{1/2}$. This means v_t and w_t have mean zero and variance-covariance matrices, $V[v_t] = V_t - \frac{2}{\pi}\Delta_t\Delta_t'$ and $V[w_t] = W_t$. This can be equivalently written as

$$x_t = G_t x_{t-1} + w_t, \tag{4.2.7}$$

$$y_t = F_t x_t + \Delta_t Z_t + u_t, \tag{4.2.8}$$

where $Z_t \sim HN_1(b, 1)$ with $b = -\sqrt{\frac{2}{\pi}}$, where HN is the half-normal distribution, which is independent of $u_t \sim N_n(0, V_t - \Delta_t\Delta_t')$ and $w_t \sim N_m(0, W_t)$. Equations (4.2.7) and (4.2.8) give the skew normal dynamic linear model. Similar to before, the filtering equations are split into two steps.

Time update equations.

$$\begin{aligned} m_{t|t-1} &= G_t m_{t-1|t-1}, \\ C_{t|t-1} &= G_t C_{t-1|t-1} G_t' + W, \\ \eta_{t|t-1} &= \eta_{t-1|t-1} B_{t-1}(1 + \eta_{t-1|t-1} H_t \eta_{t-1|t-1}')^{-1/2}, \end{aligned} \tag{4.2.9}$$

With $H_t = C_{t-1|t-1} - B_{t-1}C_{t|t-1}B'_{t-1}$ and $B_t = C_{t|t}G'_{t+1}C^{-1}_{t+1|t}$.

Measurement update equations.

$$\begin{aligned} f_t &= F'_t m_{t|t-1} - b\Delta_t, \\ Q_t &= F'_t C_{t|t-1} F_t + V, \\ m_{t|t} &= m_{t|t-1} + C_{t|t-1}F_t Q_t^{-1}(y_t - f_t), \\ C_{t|t} &= C_{t|t-1} - C_{t|t-1}F_t Q_t^{-1} F'_t C_{t|t-1}, \\ \eta_{t|t} &= \begin{pmatrix} -\Lambda_t F'_t \\ \\ \eta_{t|t-1} \end{pmatrix} + \begin{pmatrix} \Lambda_t V_t Q_t^{-1} \\ \\ C_{t|t-1}F_t Q_t - 1 \end{pmatrix} (y_t - f_t)(\theta_t - m_{t|t})^{-1}, \end{aligned} \tag{4.2.10}$$

Using properties of the multivariate normal distribution, we can prove this using induction.

Property 4.2.1. *Let $Z \sim N_k(\mu, \Sigma)$. Then for any fixed vector $\boldsymbol{u} \in \mathbb{R}^m$ and matrix $A \in \mathbb{R}^{mxk}$ (Arellano-Valle & Genton 2005),*

$$E[\Phi_m(AZ + \boldsymbol{u}|\Omega)] = \phi_m(\boldsymbol{u}|\Omega + A\Sigma A'). \tag{4.2.11}$$

Property 4.2.2. *We also make use of the marginal-conditional relation:*

$$\phi_k(x|\mu, \Sigma)\phi_m(y|\eta + A(x-\mu), \Psi) = \phi_k(x|\mu, \Sigma A'\Omega^{-1}(y-\eta), \Sigma - \Sigma A'\Omega^{-1}A\Sigma)\phi_m(y|\eta, \Omega) \tag{4.2.12}$$

The prior distribution of the state model parameters at time $t-1$ is

$$\theta_{t-1}|D_{t-1} \sim SN(m_{t-1|t-1}, C_{t-1|t-1}, \eta_{t-1|t-1}).$$

Supposing this is true, then the state model parameters at time t given the information at time $t-1$ is

$$\theta_t|D_{t-1} \sim SN(m_{t|t-1}, C_{t|t-1}, \eta_{t|t-1}).$$

Proof.

$$\begin{aligned} p(\theta_t|D_{t-1}) &= \int_{\mathbb{R}^p} p(\theta_t|\theta_{t-1})p(\theta_{t-1}|D_{t-1})d\theta_{t-1} \\ &\propto \int_{\mathbb{R}^p} \phi_p(\theta_t|G_t\theta_{t-1}, W_t)\times \\ &\qquad \phi_p(\theta_{t-1}|m_{t-1|t-1}, C_{t-1|t-1})\Phi_{t-1}(\eta_{t-1|t-1}(\theta_{t-1} - m_{t-1|t-1}))d\theta_{t-1} \end{aligned}$$

by equation 4.2.12

$$\begin{aligned} &= \phi_p(\theta_t|m_{t|t-1}, C_{t|t-1}) \int_{\mathbb{R}^p} \phi_p(\theta_{t-1} - m_{t-1|t-1}|C_{t-1|t-1}G_t'C_{t|t-1}^{-1}(\theta_t - m_{t|t-1}), \\ &\qquad C_{t-1|t-1} - C_{t-1|t-1}G_t'C_{t|t-1}^{-1}G_tC_{t-1|t-1})\Phi_{t-1}(\eta_{t-1|t-1}(\theta_{t-1} - m_{t-1|t-1}))d\theta_{t-1} \end{aligned}$$

by equation 4.2.11

$$= \phi_p(\theta_t|m_{t|t-1}, C_{t|t-1})\Phi_{t-1}(\eta_{t-1|t-1}B_{t-1}(1 + \eta_{t-1|t-1}H_t\eta_{t-1|t-1}')^{-1/2}(\theta_t - m_{t|t-1}))d\theta_t$$

where $\eta_{t|t-1} = \eta_{t-1|t-1}B_{t-1}(1 + \eta_{t-1|t-1}H_t\eta_{t-1|t-1}')^{-1/2}$. Thus, this gives the kernel of the pdf of $\theta_t|D_{t-1} \sim SN(m_{t|t-1}, C_{t|t-1}, \eta_{t|t-1})$. □

The one-step forecast distribution at time t given the information at $t-1$ is

$$Y_t|D_{t-1} \sim N(f_t, Q_t).$$

Proof.

$$\begin{aligned} p(y_t|D_{t-1}) &= \int_{\mathbb{R}^p} p(y_t|\theta_t, D_{t-1})p(\theta_t|D_{t-1})d\theta_t \\ &\propto \int_{\mathbb{R}^p} \phi_r(y_t|F_t'\theta_t - \sqrt{\frac{2}{\pi}}, V_t)\Phi_1(\Lambda_t(y_t - F_t'\theta_t - \sqrt{\frac{2}{\pi}})) \\ &\qquad \times \phi_p(\theta_t|m_{t|t-1}, C_{t|t-1})\Phi_{t-1}(\eta_{t|t-1}(\theta_t - m_{t|t-1}))d\theta_t \end{aligned}$$

by equation 4.2.12

$$= \phi_r(y_t|f_t, Q_t) \int_{\mathbb{R}^p} \phi_p(\theta_t|m_{t|t}, C_{t|t}) \Phi_t(\eta_{t|t}(\theta_t - m_{t|t}))d\theta_t$$

where

$$\eta_{t|t} = \begin{pmatrix} -\Lambda_t F_t' \\ \\ \eta_{t|t-1} \end{pmatrix} + \begin{pmatrix} \Lambda_t V_t Q_t^{-1} \\ \\ C_{t|t-1} F_t Q_t - 1 \end{pmatrix} (y_t - f_t)(\theta_t - m_{t|t})^{-1}. \tag{4.2.13}$$

Now, by equation 4.2.11 we have

$$= \phi_r(y_t|f_t, Q_t) \Phi_t(0|\eta_{t|t} C_{t|t} \eta_{t|t}' + I),$$

where I is the identity matrix. Since $\Phi_t(0|\eta_{t|t} C_{t|t} \eta_{t|t}' + I) = 1/2$ we have

$$p(y_t|D_{t-1}) \propto \phi_r(y_t|f_t, Q_t),$$

which is the kernel of the pdf of $y_t|D_{t-1} \sim N(f_t, Q_t)$. □

The state parameters at time t given the information at time t is given by

$$\theta_t|D_t \sim SN(m_{t|t}, C_{t|t}, \eta_{t|t}).$$

Proof. Using the relations $\Lambda_t(y_t - F_t'\theta_t + \sqrt{\frac{2}{p\imath}}\Delta) = -\Lambda_t F_t'(\theta_t - m_{t|t}) + \Lambda_t V_t Q_t^{-1}(y_t - f_t)$

and $m_{t|t} - m_{t|t-1} = C_{t|t-1}F_tQ_t^{-1}(y_t - f_t)$. Similar to before

$$\begin{aligned}
p(\theta_t|y_t, D_{t-1}) \propto & p(y_t|\theta_t, D_{t-1})p(\theta_t|D_{t-1}) \\
\propto & \phi_r(y_t|F_t'\theta_t - \sqrt{\frac{2}{\pi}}, V_t)\times \\
& \Phi_1(\Lambda_t(y_t - F_t'\theta_t - \sqrt{\frac{2}{\pi}}))\phi_p(\theta_t|m_{t|t-1}, C_{t|t-1})\Phi_{t-1}(\eta_{t|t-1}(\theta_t - m_{t|t-1})) \\
\propto & \phi_p(\theta_t|m_{t|t}, C_{t|t})\Phi_1(\Lambda_t(y_t - F_t'\theta_t - \sqrt{\frac{2}{\pi}}))\Phi_{t-1}(\eta_{t|t-1}(\theta_t - m_{t|t-1})) \\
= & \phi_p(\theta_t|m_{t|t}, C_{t|t})\Phi_t\left(\begin{pmatrix} -\Lambda_t F_t' \\ \eta_{t|t-1} \end{pmatrix}(\theta_t - m_{t|t}) + \begin{pmatrix} \Lambda_t V_t Q_t^{-1} \\ C_{t|t-1}F_tQ_t - 1 \end{pmatrix}(y_t - f_t)\right)
\end{aligned}$$

which we can rewrite as

$$\begin{aligned}
& = \phi_p(\theta_t|m_{t|t}, C_{t|t})\Phi_t\left(\left(\begin{pmatrix} -\Lambda_t F_t' \\ \eta_{t|t-1} \end{pmatrix} + \begin{pmatrix} \Lambda_t V_t Q_t^{-1} \\ C_{t|t-1}F_tQ_t - 1 \end{pmatrix}(y_t - f_t)(\theta_t - m_{t|t})^{-1}\right)(\theta_t - m_{t|t})\right) \\
& = \phi_p(\theta_t|m_{t|t}, C_{t|t})\Phi_t(\eta_{t|t}(\theta_t - m_{t|t}))
\end{aligned}$$

Thus, this gives the kernel of the pdf of $\theta_t|D_t \sim SN(m_{t|t}, C_{t|t}, \eta_{t|t})$, concluding the proof. □

As a result of deriving the Kalman filter for the skew normal distribution, the latent variable is present in $\eta_{t|t}$. However, this should not be an issue as η does not appear in the filtering equations for m or C, nor the likelihood.

Example 4.2.2. (Skew Kalman filter for a Gaussian random walk with skewed noise) Now, applying the skew Kalman filter to the observations y_t, given by the Gaussian random walk in Example 4.2.1, we can estimate the state x_t. Figure 4.2.4 shows the Skew Kalman filter applied to the Gaussian random walk. Compared to the traditional filter, shown in Figure 4.2.2, the skew filter is able to correct for the skew present and visually captures the underlying signal better than the traditional

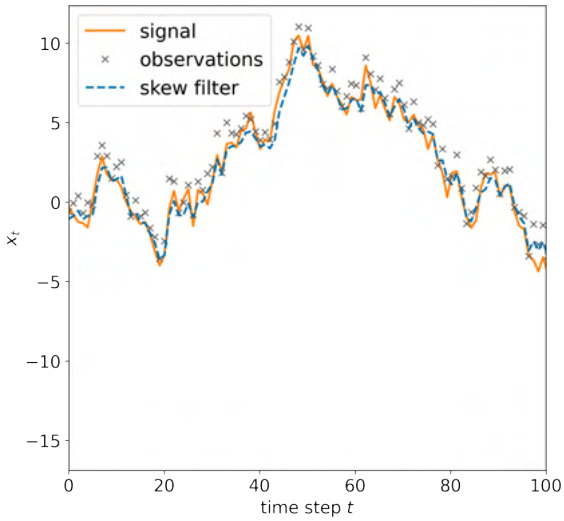


Figure 4.2.4: Skew Kalman filter applied to the Gaussian random walk from Example 4.2.1 with parameters W=V=1 and $\lambda = 2$.

Kalman filter.

4.2.3 Parameter Estimation

Following our notation from Chapter 3, the likelihood for the basic filter is given by

$$\mathcal{L} = \prod_{t=1}^{n} \phi(y_t | m_{t,t-1}, P_{t,t-1} + S), \tag{4.2.14}$$

where S is the observation noise parameter, $m_{t,t-1}$ and $P_{t,t-t}$ are the estimated state and covariance at time t. For the skew normal Kalman filter, we need to estimate the noise parameters W and V as well as the skew parameter λ. From our proof of the one-step forecast distribution we have that the likelihood for the skew normal Kalman filter is given by

$$\mathcal{L} = \prod_{t=1}^{n} \phi(y_t | F_t' m_{t,t-1} - b\Delta_t, F_t' C_{t|t-1} F_t + V), \tag{4.2.15}$$

where V is the observation noise parameter and $m_{t,t-1}$ and $C_{t,t-t}$ are the estimated state and covariance at time t, F is the observation matrix, and $b = -\sqrt{\frac{2}{\pi}}$. The

additional skew information is given by $\Delta_t = V_t^{1/2}\delta_t$, where $\delta_t = \lambda_t/(1+\lambda_t'\lambda_t)^{1/2}$ and λ is the skewness parameter. This is no more complicated than the basic Kalman filter likelihood. Thus, this preserves the computational efficiency and simplicity of the basic filter while still incorporating the additional skew information. Furthermore, compared to the unified skew normal pdf given in Equation (4.2.3), this is more computationally tractable as we have avoided the high dimensional matrices in the cdf.

Likelihood plots for W, V and λ are shown in Figure 4.2.5. While we can easily estimate W using this likelihood, V and λ are linked parameters and more difficult to estimate. From the right most plots we can see that the range over which the likelihood varies for different values of λ is very small. Therefore, W and V will produce the same estimates regardless of the value of λ. If the true parameters are $W = V = 1$ and $\lambda = 0$, if λ is fixed at any value it will still estimate the noise parameters W and V close to 1.

If $Y \sim SN(\eta, \Omega, \lambda)$ then we can rewrite the likelihood for the skew Kalman filter as a skew normal using

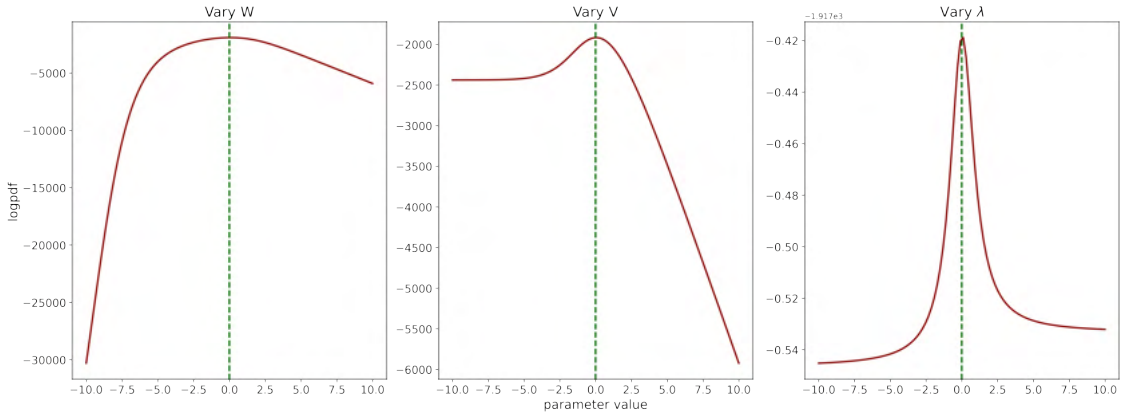
$$Y = \eta + \Delta Z + U \tag{4.2.16}$$

where $\Delta = \Omega^{1/2}\delta$, Z is a half normal $Z \sim N(0,1)$, and $\delta = \lambda/\sqrt{1+\lambda^2}$. We find that given $\eta = f_t$ and $\Omega = Q_t(1-2/\pi\delta^2)^{1/2}$ we can rewrite our normal likelihood as a skew normal given by

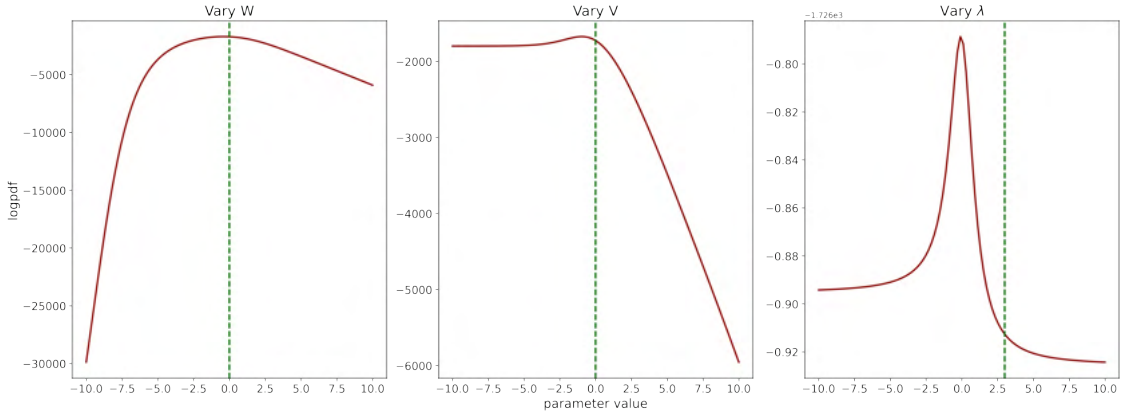
$$y_t \sim SN(f_t, Q_t(1-\frac{2}{\pi}\delta^2)^{-1}, \lambda), \tag{4.2.17}$$

if the skewness parameter $\lambda = 0$ we have $y_t \sim SN(f_t, Q_t, 0)$, which is equivalent to the Gaussian likelihood, $y_t \sim N(f_t, Q_t)$ as expected.

Likelihood plots for the new likelihood are shown in Figure 4.2.6. While this



(a) Data generated using W = 1, V = 1, $\lambda = 0$.



(b) Data generated using W = 1, V = 1, $\lambda = 3$.

Figure 4.2.5: Log likelihood plots for the skew Kalman filter using the log of the likelihood given in Equation (4.2.15). 1000 data points were simulated using $W = V = 1$ and **(a)** $\lambda = 0$ and **(b)** $\lambda = 3$. The green dashed lines show the true values, and the optimised result for the estimated parameter is given in the titles of each plot. We optimate over the exponential of the noise parameters so we can estimate them over the real line. Therefore, data generated with $W = V = 1$ corresponds to a true value of 0.

has addressed the short range over which the likelihood varies for changing λ, the parameter λ still always estimates close to 0. The relationship between the basic filter parameters and the skew filter parameters is as follows:

$$R = W, \tag{4.2.18}$$

$$S = V(1 - \frac{2\delta^2}{\pi}) \qquad \text{with} \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}. \tag{4.2.19}$$

Thus, when $\lambda = 0$ we have that S=V, which is simply the standard Kalman filter. Due to the relationship between V and λ, V is also being incorrectly estimated, if λ were to be fixed at the correct value V would also be correct. Thus, we seek an alternative method in which to estimate the skewness parameter λ.

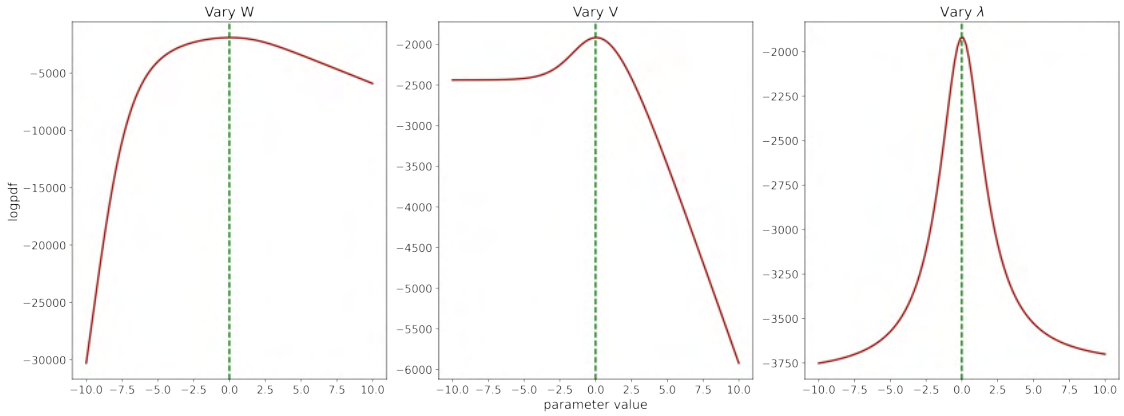
Since there is only one parameter to estimate this can be done using a grid search over possible parameters and choosing the parameter that minimises some chosen metric. Here we will use the mean square error between the filter estimate and underlying signal as this is known in the simulation study we carry out below.

The steps for obtaining the skew Kalman filter estimate are as follows:

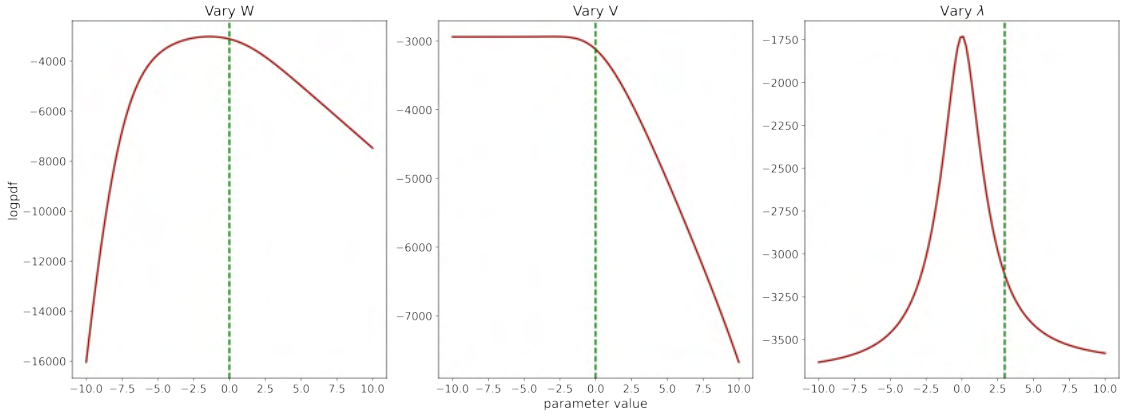
1. Run the skew Kalman filter with $\lambda = 0$ to estimate W and V.
2. Carry out a grid search over values of λ, updating V according to Equation 4.2.19, to find the value of λ that minimises MSE.
3. Obtain skew filter estimate using new parameters.

4.3 Simulation Study

In this section we consider a simulation study to evaluate the performance of the skew normal Kalman filter and the proposed method to estimate the skew Kalman



(a) Data generated using W = 1, V = 1, $\lambda = 0$.



(b) Data generated using W = 1, V = 1, $\lambda = 3$.

Figure 4.2.6: Log likelihood plots for the skew Kalman filter using the log of the likelihood given in Equation (4.2.17). 1000 data points were simulated using $W = V = 1$ and **(a)** $\lambda = 0$ and **(b)** $\lambda = 3$. The green dashed lines show the true values and he optimised result for the estimated parameter is given in the titles of each plot. We optimate over the exponential of the noise parameters so we can estimate them over the real line. Therefore, data generated with $W = V = 1$ corresponds to a true value of 0.

filter parameters. For this we consider the following skew normal DLM

$$\theta_t = \theta_{t-1} + w_t, \tag{4.3.1}$$

$$Y_t = \theta_t + v_t, \tag{4.3.2}$$

with $v_t \sim SN(0, V, \lambda)$ and $w_t \sim (0, W)$, for $t = 1, ..., 1000$. We compare the performance of the skew Kalman filter to the traditional Kalman filter both in the presence of skew and when there is no skew. In the absence of skewness, the skew Kalman filter reduces to the traditional Kalman filter. Equations (4.2.18) and (4.2.19) are used to calculate the true parameters for the traditional Kalman filter. We consider 3 cases: $\lambda = 0$, to ensure the skew filter performs the same as the traditional Kalman filter when there is no skew, a positive skew case where $\lambda = 2$, and a negative skew case where $\lambda = -0.5$. As by the proposed steps for estimating parameters using the skew Kalman filter the underlying process is used to estimate the skewness parameter and minimise the error between the 2 datasets. The traditional Kalman filter is not using the information from the underlying process and is estimating the parameters from the observation data only.

4.3.1 Case 1: $\lambda = 0$

First, we simulate a signal using Equation (4.3.1) and $W = 1$. Second, we simulate 100 series of measurements from Equation (4.3.2) with $V = 1$ and $\lambda = 0$.

Figure 4.3.1 shows the filter results for both the basic and skew filter using the mean parameter estimates. Both the basic filter and the skew filter perform similarly, which is to be expected in the absence of any skew in the data as the skew Kalman filter reduces to the basic filter. The MSE for the basic filter and skew filter is 0.60 for both, which further confirms they are performing similarly. Table 4.3.1 summarises the MLE results for the SKF and the basic KF. The parameter estimates for each filter are reasonably close to the true value, with the skew Kalman filter

Table 4.3.1: MLE results based on 100 generated data sets for $W = V = 1$ and $\lambda = 0$.

Parameter	True Value	Mean Estimate	SD Estimate
W	1	0.955	0.080
V	1	1.001	0.105
λ	0	0.045	0.059
R	1	0.970	0.080
S	1	0.997	0.105

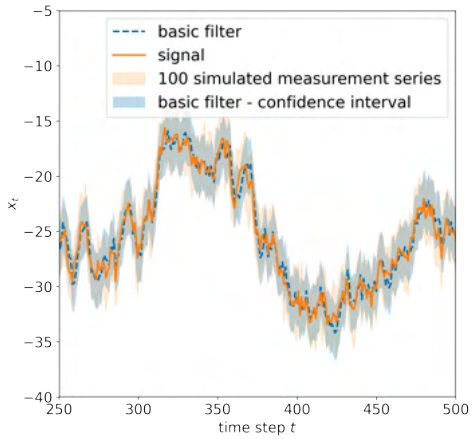
Mean and standard deviation of the parameter estimates for the 100 simulated data sets. Upper box shows the parameter estimates for the skew normal Kalman filter and the lower box shows the parameter estimates for the traditional Kalman filter.

mean estimate being between 0.001 and 0.045 of the true value and the traditional Kalman filter being between 0.003 and 0.3 of the true value.

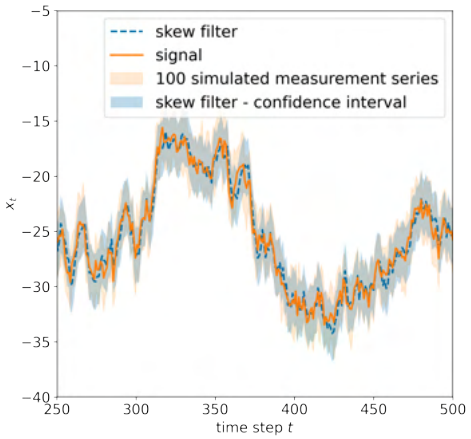
4.3.2 Case 2: $\lambda = 2$

We repeat the same experiment again but with $\lambda = 2$.

Figure 4.3.2 shows the filter results for both the basic and skew filter using the mean parameter estimates. In Figure 4.3.2a, the basic filter estimate is consistently higher than the true underlying signal. This is because the basic filter is unable to capture the skew in the data and returns a state estimate which closely represents the mean of the data rather than the underlying signal. Comparing this to the skew filter estimate in Figure 4.3.2b, since the skew filter is able to adjust for the skew in the data it captures the underlying signal better. This results in a lower MSE for the SKF of 0.60, compared to the standard KF where it is 0.89. Table 4.3.2 summarises the MLE results for the SKF and the basic KF. The true parameter values for the



(a) Filter results for basic filter.

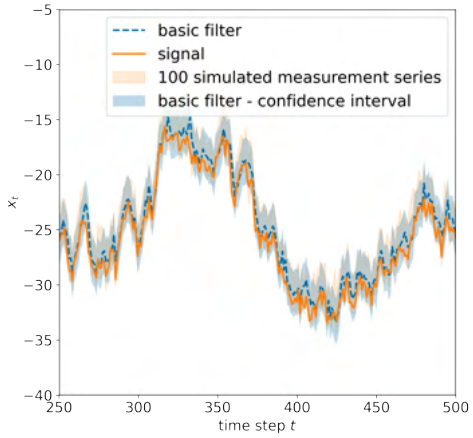


(b) Filter results for skew filter.

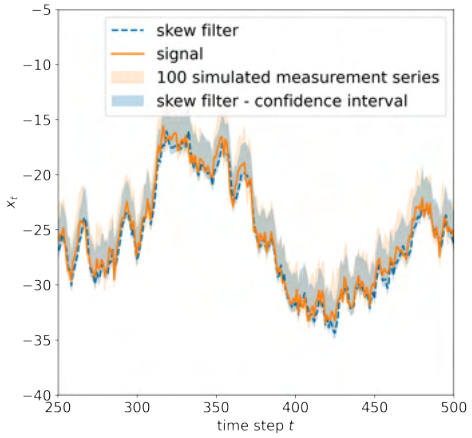
Figure 4.3.1: Filter results for basic and skew filter (dashed, blue line). Simulated signal using Equation (4.3.1) and $W = 1$ (solid, orange line). 100 simulated measurement series from Equation (4.3.2) with $V = 1$ and $\lambda = 0$ (shaded region, yellow) and the mean of the simulated measurement series at each time (dashed line, grey).

Table 4.3.2: MLE results based on 100 generated data sets for $W = V = 1$ and $\lambda = 2$. Mean and standard deviation of the parameter estimates for the 100 simulated data sets. Upper box shows the parameter estimates for the skew normal Kalman filter and the lower box shows the parameter estimates for the basic Kalman filter.

Parameter	True Value	Mean Estimate	SD Estimate
W	1	0.937	0.055
V	1	0.980	0.073
λ	2	1.78	0.322
R	1	0.950	0.055
S	0.491	0.509	0.065



(a) Filter results for basic filter with simulated data.



(b) Filter results for skew filter with simulated data.

Figure 4.3.2: Filter results for basic and skew filter (dashed, blue line). Simulated signal using Equation (4.3.1) and $W = 1$ (solid, orange line). 100 simulated measurement series from Equation (4.3.2) with $V = 1$ and $\lambda = 2$ (shaded region, yellow) and the mean of the simulated measurement series at each time (dashed line, grey).

basic KF are adjusted using Equations 4.2.18 and 4.2.19. The mean estimated skew is slightly lower than the true value at 1.78 with a std of 0.3. However, this has not had a large impact on the estimate of V and the SKF still performs better compared to the basic KF.

4.3.3 Case 3: $\lambda = -0.5$

Finally, we repeat the same experiment again but with $\lambda = -0.5$. Figure 4.3.3 shows the filter results for both the basic and skew filter using the mean parameter estimates. Now, due to the negative skew, the traditional Kalman filter estimate is consistently lower than the underlying signal, whereas the skew Kalman filter is able to adjust for this. Again, this results in a lower MSE for the SKF, 0.60, compared to the standard KF, 0.68. Table 4.3.3 summarises the MLE results for the SKF and the basic KF. Again, true parameter values for the basic KF are adjusted using

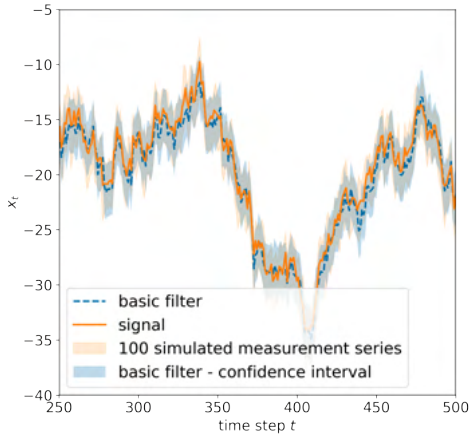
Table 4.3.3: MLE results based on 100 generated data sets for $W = V = 1$ and $\lambda = -0.5$. Mean and standard deviation of the parameter estimates for the 100 simulated data sets. Upper box shows the parameter estimates for the skew normal Kalman filter and the lower box shows the parameter estimates for the basic Kalman filter.

Parameter	True Value	Mean Estimate	SD Estimate
W	1	0.969	0.073
V	1	1.026	0.113
λ	-0.5	-0.43	0.076
R	1	0.975	0.072
S	0.873	0.924	0.072

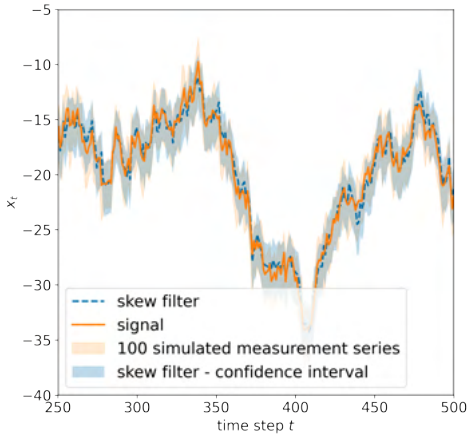
Equations 4.2.18 and 4.2.19. The mean parameter estimates are close to the true value for each of the parameters for both filters.

4.4 Discussion and Conclusion

The simulation study demonstrates that correctly estimating the skewness allows for a better capturing of the underlying signal and a reduction in error. We can see from Cases 2 and 3 that a higher value in skew leads to a greater MSE in the basic KF, whereas the SKF performs similarly across all the cases. Thus, the higher values in skewness lead to a greater improvement using the SKF compared to the basic KF. To evaluate the suitability for higher values of skewness we look at the estimated skewness over a range of values. Similar to the simulation study, 25 sets of noisy measurements were simulated for the same underlying signal. The true value of skew is increased at equal intervals from -5 to 5 and we consider the mean estimate of the skew from the 25 simulated measurement sets. From Figure 4.4.1 we can see the estimates perform best when close to 0 and as you move away from



(a) Filter results for basic filter with simulated data.



(b) Filter results for skew filter with simulated data.

Figure 4.3.3: Filter results for basic and skew filter (dashed, blue line). Simulated signal using Equation (4.3.1) and $W = 1$ (solid, orange line). 100 simulated measurement series from Equation (4.3.2) with $V = 1$ and $\lambda = -0.5$ (shaded region, yellow) and the mean of the simulated measurement series at each time (dashed line, grey).

zero the skew estimate moves further away from the true value. While the best performance is seen between $(-2, 2)$, even with the estimated skew being too low for the higher values of skewness, there will still be an improvement compared to the basic KF.

The main limitation of the simplified skew Kalman filter is that estimating the parameters using the 2-step method proposed here requires the underlying process to be known to be able to estimate the skew parameter. While knowledge of the underlying signal is limiting in terms of applications, the simplified skew Kalman filter would be suitable for bias correction, which we consider in Chapter 5. If the skew was relatively constant, the underlying process may only be needed for a training portion of the data. However, in the absence of anything to indicate the skew or some informed assumptions as to what it should be, the skew will always estimate close to 0.

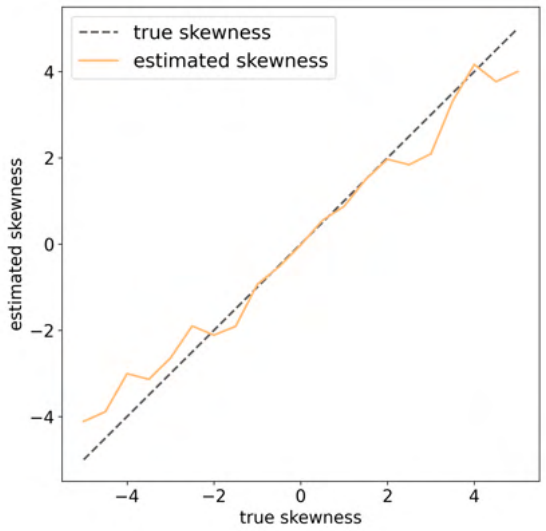


Figure 4.4.1: Mean estimate of the skewness parameter (solid line), from 25 simulated measurement sets, and the true value (dashed line) using a grid search to estimate the parameters.

Recent work by Wang et al. (2023), which shares authors with the paper which motivated this research (Arellano-Valle et al. 2019), has explored the identifiability issue with the unified skew normal distribution. They propose to constraining parameters to enforce identifiability. One possible option being to limit $\tau = 0$ which corresponds with our initial investigation of the distribution. Further, by limiting τ and reducing the skew Kalman filter to the multivariate skew normal we are able to avoid the computational problems in the unified skew normal Kalman filter caused by large matrices in the SUN pdf and successfully estimate the parameters using a combination of MLE and a grid search, all of which preserves the simplicity of the standard Kalman filter. Thus, we have successfully implemented a simplified version of the skew Kalman filter, that is still relatively simple and computationally efficient.

Chapter 5

A Skew Kalman Filter Approach for Bias Correction and Infilling Missing Data, Demonstrated for Surface Ozone

Abstract

Incomplete data is a common issue in air quality monitoring, and often results from instrument failure or maintenance. Reanalysis data offers a time complete picture of air quality providing an attractive source to infill gaps in the observation data. Since it is often biased, using 'raw' reanalysis data to infill missing data is likely to introduce errors into any statistical inference based on the in-filled dataset. Here, we propose a method for infilling missing data that considers the bias between the datasets as a skewness allowing for the use of a skew Kalman filter, an extension to the traditional filter that allows for skewness to be present in the observation noise. The Kalman filter is a simple, computationally efficient tool for estimating the underlying process of a system under noisy measurements and provides an associated uncertainty in the estimate. For our real-world missing data scenario using surface

level ozone, the reduction in RMSE was 0.1 ppb - 1.3 ppb across 8 measurement sites compared to 'raw' reanalysis data. By resolving the bias issue, we can implement a computationally efficient statistical method for an improved way in which to infill the missing data that is transferable to other applications.

5.1 Introduction

Environmental time series from measurement sites often contain missing data due to events like instrument failures, power cuts or instrument calibration issues. At the same time, we often need a complete time series from a given location to inform site-specific inference (e.g., crop damage or health impacts from pollution). While there are lots of infilling methods, (e.g. linear interpolation, regression based methods (Junninen et al. 2004) or neural networks (Zhang & Thorburn 2022)), each have advantages as well as drawbacks. Here, we describe a novel approach for infilling missing data using a skew Kalman filter (Naveau et al. 2005), which we demonstrate with measurement time series of the air pollutant ozone. This approach has the advantage of being computationally efficient, relatively straight forward to implement and able to produce uncertainty estimates for the infilled data.

Methods for infilling missing data range from simple to complex (Hartley & Hocking 1971, Little & Rubin 2002, and references therein). Simple methods include linear interpolation, which connects the end points of the gap with a straight line, infilling with the mean or median of non-missing observations, and case deletion, where missing data is simply ignored. While straight forward to use, these methods can disrupt the structure of the data and introduce large errors into the analysis (Baraldi & Enders 2010, Donders et al. 2006). Other more sophisticated approaches include regression-based imputation methods, which are based on estimated regression models between missing and available data (Mirzaei et al. 2022); the expectation-maximisation method, which is an iterative method for estimating missing data

(Schneider 2001, Baraldi & Enders 2010); neural networks, which are inspired by biological neural networks and can capture and make use of the temporal information, as well as computing other correlated predictors to infill missing data (Choudhury & Pal 2019, Coulibaly & Evora 2007); self organizing maps, which are a tool for multivariate data mapping data from a high dimensional input space to a low dimensional output space, serving as a clustering tool that can interpolate between previously encountered inputs (Lamrini et al. 2011, Folguera et al. 2015, Nkiaka et al. 2016); and singular spectrum analysis, which is a non-parametric technique for time series analysis that aims to reproduce the original time series from its principle components (Shen et al. 2015, Schoellhamer 2001). However, these methods can be computationally expensive or require additional covariate information, particularly if modelling the data is required.

Aside from infilling the data based on covariates, we could alternatively make use of output from physically based process models, which provide a spatiotemporally complete picture of the measured variable. Reanalysis data is a data assimilation product that assimilates multiple observations into a physically based process model of the atmosphere to give a best estimate of the state of the atmosphere, including for physical parameters not measured or not measurable (Kanamitsu et al. 2002, Dee et al. 2014, Hersbach et al. 2020). Despite being calibrated to observational data, reanalyses still contain biases, particularly when compared to local scale measurement data (Casciaro et al. 2022, Wagner et al. 2020).

Here, we demonstrate a method that uses reanalysis output to infill missing data from measurement sites, while also correcting the bias between the two datasets. Our approach conceptualises the bias as a 'skewness', between the two datasets, which facilitates the use of an extension to the standard Kalman filter to correct the bias. The standard Kalman filter itself is a useful tool for analysing time series data and continues to be well used across various disciplines long after it was first

proposed by Kalman (1960). Examples of work implementing the Kalman filter alongside air pollution data fall into three main categories: 1) post processing, where a Kalman filter is applied to climate model forecasts to reduce the errors and improve the accuracy of the forecasts (Djalalova et al. 2015, Heemink & Segers 2002, Ridder et al. 2012); 2) inverse modelling methods, where Kalman filters are used to identify biases and improve emission inventories using satellite observations (Napelenok et al. 2008, Metia et al. 2020); and 3) producing forecasts, which use Kalman filters with observational data and covariates, such as meteorological effects and human activity (Hoi et al. 2008, Zolghadri & Cazaurang 2006). While the standard Kalman filter assumes normality in the observation noise, the skew Kalman filter allows for greater flexibility in the noise distribution, through inclusion of an additional skewness parameter. This extends the applicability of the Kalman filter to a wider range of datasets without compromising the low-cost computational benefits (Arellano-Valle et al. 2019, Fasano et al. 2019). Here, we use a skew Kalman filter that follows a multivariate skew normal distribution, allowing information from the skew between two datasets to be included in the filter estimate. Section 5.2 further explains the use of the skew between the datasets as a bias correction technique.

We demonstrate our method using surface ozone data, for which reanalysis and measurement data are freely available. Ozone is an air pollutant that has negative impacts on human health (Huangfu & Atkinson 2020) and ecosystems (Wittig et al. 2009), as well as crop yields (Feng et al. 2008). The analysis of surface level observations is vital to improving the understanding of the spatiotemporal distribution of ozone and its trends (Tarasick et al. 2019). The recent Tropospheric Ozone Assessment Report (TOAR) collated surface ozone from $\sim$ 10,000 monitoring stations globally (Schultz et al. 2017). Like other environmental measurement data, these surface ozone measurements are often incomplete, which can present issues for analysing and characterising the data as well as utilising it as an input to site-level models. Here, we demonstrate and evaluate our method for the

years 2015-2017, inclusive, using 15 sites in Germany from the TOAR database and reanalysis data from the Copernicus Atmospheric Modelling System (CAMS; https://atmosphere.copernicus.eu).

The rest of this paper is organised as follows. In Section 5.2, we describe our skew Kalman filter approach to infilling and bias correction and introduce CAMS and TOAR data. In Section 5.3, we assess the ability of our skew Kalman method for correcting the bias in the reanalysis data and as a method for infilling randomly missing data. Finally, in Section 5.4 we discuss the comparative advantages and potential extensions to our approach.

5.2 Methods and Data

5.2.1 Skew Normal Kalman Filter

The bias between the two datasets can be considered as an unknown skew term, where a positive skew arises from an overestimate from the biased dataset and conversely, a negative skew arises from an underestimate. A skewness of zero implies no skewness, and thus no bias. The skew normal Kalman filter (SKF) (Naveau et al. 2005) is an extension of the Kalman filter (KF), which is a computationally efficient recursive algorithm for tracking a time-dependent state vector in real time, θ_t, with a noisy evolution equation and noisy measurements, y_t (Kalman 1960). A KF seeks to estimate the true underlying process of a system under noisy measurements. The SKF extends the assumption of normality on the observation noise to include skew normal distributed observations, with state and measurement equations are given by

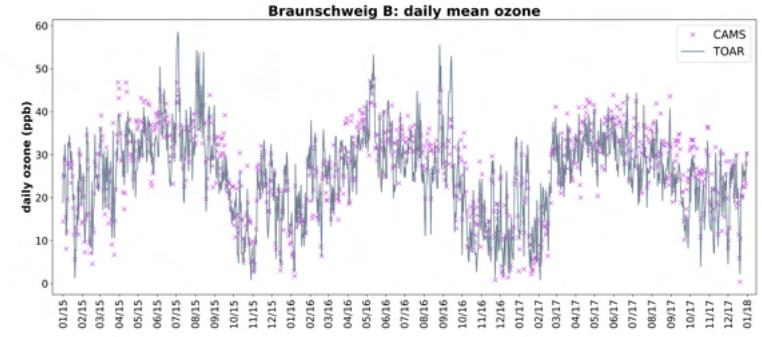
$$\theta_t = G_t\theta_{t-1} + w_t, \tag{5.2.1}$$

$$y_t = F_t\theta_t + v_t, \tag{5.2.2}$$

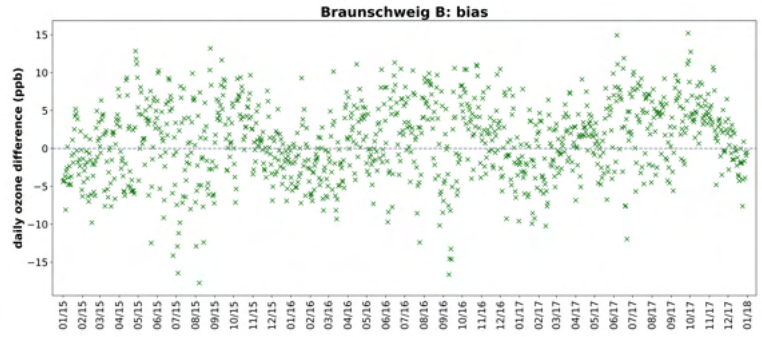
respectively, where the process noise, $w_t \sim N_n(0, W)$, is normally distributed and $v_t \sim SN_n(0, V, \lambda)$ is a skew normal distribution with skew parameter λ. Matrices W and V are positive semi-definite covariance matrices. The matrix G_t is the transition matrix and F_t is the measurement model matrix.

Figure 5.2.1a compares the reanalysis data from CAMS and the surface observation data from TOAR for a suburban background site located in Broitzem, Braunschweig 2015 to 2017. By treating the reanalysis data as noisy measurements of some underlying system, assumed to be represented by the observation data, the resulting Kalman filter estimate should be closer to the observation data than the reanalysis data itself. Figure 5.2.1b shows the difference between the CAMS and TOAR concentrations, the mean difference between the datasets is 1.0, as this is greater than 0, this indicates there is a bias present and the reanalysis data is overestimating the ozone concentrations compared to TOAR. Across the sites that will be discussed is this paper the mean difference between the reanalysis data and the observation data ranges between 0.2 and 2.7. Thus, the reanalysis ozone concentrations are not symmetric about the observation ozone concentrations. The skew normal distribution is not symmetric about the mean and thus is a more suitable approximation for the noise present in this system.

Current literature for the SKF uses an extension to the multivariate skew normal, the univariate skew normal (SUN) (Arellano-Valle & Azzalini 2006) or closed skew normal. This has been the preferred choice since the univariate skew normal preserves the conditional and marginal distributions (Arellano-Valle & Azzalini 2020) whilst the multivariate skew normal family preserves only the marginal distribution. However, these distributions can be computationally intractable making it difficult to estimate the fixed parameters (Wang et al. 2023). Here, we adopt a simplified approach to the SKF, as proposed by Arellano-Valle et al. (2019), which reduces the unified skew normal to a standard skew normal. We then



(a)



(b)

Figure 5.2.1: **(a)** CAMS and TOAR daily mean ozone and **(b)** the difference in ozone concentrations (ppb) between CAMS and TOAR at Braunschweig B during 2015 to 2017.

implement a two-step approach to estimate the parameters. The filtering equations for the SKF can be written as two steps, the time update and the measurement update.

We assume an initial skew normal distribution for θ_0, at time $t = 0$, with mean m_0 and covariance C_0. Using equations 5.2.3 and 5.2.4 we can recursively update the mean and variance of θ_t. We use the time update equations to obtain a mean and covariance at time t using information up to time $t-1$, denoted by $t|t-1$:

Time update equations

$$\begin{aligned} m_{t|t-1} &= G_t m_{t-1|t-1}, \\ C_{t|t-1} &= G_t C_{t-1|t-1} G_t' + W. \end{aligned} \tag{5.2.3}$$

The measurement update equations are then used to obtain a mean and covariance at time t using the information up to time t, including the measurement at time t, denoted by $t|t$:

Measurement update equations

$$\begin{aligned} f_t &= F_t' m_{t|t-1} - b\Delta_t, \\ Q_t &= F_t' C_{t|t-1} F_t + V, \\ m_{t|t} &= m_{t|t-1} + C_{t|t-1} F_t Q_t^{-1} (y_t - f_t), \\ C_{t|t} &= C_{t|t-1} - C_{t|t-1} F_t Q_t^{-1} F_t' C_{t|t-1}, \end{aligned} \tag{5.2.4}$$

where $b = -(2/\pi)^{1/2}$, $\Delta_t = V_t \eta_t (1 + \eta_t' V_t \eta_t)^{-1/2} = V_t^{1/2} \delta_t$, $\eta_t = V_t^{-1/2} \lambda_t$, and $\delta_t = \lambda_t (1 + \lambda_t' \lambda_t)^{-1/2}$. These parameters result from the reparametrisation used by Arellano-Valle et al. (2019).

Compared to the KF we now have an additional term in the measurement update equations, $-b\Delta_t$, which acts as a correction term for the skew present in the measurement noise v_t, shifting the filter estimate to account for the skewness. The likelihood for the SKF is given by

$$\mathcal{L}(\psi) = \prod_{t=1}^{n} \frac{2}{\omega}\phi\left(\frac{y_t - f_t}{\omega}\right)\Phi\left(\lambda\left(\frac{y_t - f_t}{\omega}\right)\right), \tag{5.2.5}$$

where y_t denotes the observations, f_t is the location parameter and ω is the scale parameter $Q_t(1 - (2/\pi)\delta^2)^{-1}$. Here, $\phi(\cdot)$ and $\Phi(\cdot)$ represent a normal probability density function (pdf) and cumulative density function (cdf), respectively. Our parameters of interest are $\psi = (W, V, \lambda)$.

While using maximum likelihood estimation would be a relatively efficient way to estimate parameters, a limitation for this model is that the skew parameter cannot be correctly estimated with that method as only the biased dataset is used in the likelihood function. Thus, a two-step process is needed to first estimate the noise parameters, followed by the skewness parameter separately.

The estimate of the process noise W is independent of the value of the skewness parameter. However, V changes with respect to λ according to the relationship

$$V_{\lambda=0} = V_{\lambda=\alpha}\left(1 - \frac{2\delta^2}{\pi}\right) \qquad \text{with} \quad \delta = \frac{\lambda}{\sqrt{1+\lambda^2}}, \tag{5.2.6}$$

where α is some unknown constant. To estimate λ we seek to minimise the mean square error (MSE) between the biased and unbiased datasets. We do this by a grid search over λ, updating V using Equation (5.2.6), and choosing the value of λ that minimises the MSE between the datasets. From Equation (5.2.6), δ follows a logistic curve, meaning that $-1 \leq \delta \leq 1$ for large λ. Hence, we constrain λ between -10 and 10, which is sufficient to reach the upper and lower limit of δ and allowing

larger λ offers no additional information to the model.

Similarly to the traditional filter, the skew filter estimate, $m_{t|t}$, has an associated variance, $C_{t|t}$, which can be used to calculate confidence intervals for the estimate. However, while the traditional filter has a symmetric confidence interval, the skew filter confidence interval does not. This is due to the skew normal distribution not being symmetric about the mean, we calculate the 95% confidence interval numerically using the *SciPy* package in Python (Virtanen et al. 2020).

5.2.2 Data Description

The observational data used in this study consist of measured surface ozone data from the TOAR database (Schultz et al. 2017). This is the world's largest database of hourly surface ozone observations, combining data from over 10,000 measurement sites with particularly high spatial coverage in Europe, North America, South Korea and Japan (Schultz et al. 2017). The database contains a mixture of site types, including background urban, rural, suburban, roadside and industrial sites, and the measurement data are used for policy evaluation and determining compliance with pollutant safe levels. Spatiotemporally reanalysis surface ozone data was taken from the CAMS reanalysis dataset. CAMS ozone data is produced by the European Centre for Medium-Range Weather Forecasting using a 4D-Var data assimilation system, which combines ozone data from satellite retrievals with a process-based model of atmospheric physics and chemistry (Inness et al. 2019, 2015). Three-hourly CAMS data is currently available for the period January 2003 to December 2021 with a horizontal resolution of 80km.

We demonstrate our infilling method using daily mean surface ozone levels. We consider only TOAR sites with a temporally complete set of data such that we can subsequently compare our infilled data against the true value. As no sites have a complete hourly record, we choose sites for which a minimum of two thirds of the

hourly observations are present for each 24-hour period, starting from 00:00. Finally, we restrict ourselves to sites where the root mean squared error (RMSE) between the measurements and CAMS is less than 6 ppb (parts per billion); i.e., sites where the CAMS data is representative of the TOAR data. The CAMS data was sampled from the nearest grid square to the TOAR data and converted to TOAR's mixing ratio units (multiplying by the molecular weight of ozone divided by the molecular weight of air).

Following our site selection criteria, we used 15 TOAR sites in Germany, of which there are 4 urban, 5 rural, and 6 suburban sites. The locations of these sites are shown in Figure (5.2.2). Two of the sites have measurements in different networks (Stendal Stadstee and Braunschweig) and treat the members of each pair as different sites, suffixing them with A or B if they are on the AIRBASE or UBA network, respectively. The AIRBASE network (https://www.eea.europa.eu/data-and-maps) is the European air quality database maintained by the European Environment Agency and the German Environment Agency, Umweltbundesamt (UBA; https://www.umweltbundesamt.de/en/data), provides data about the concentration of pollutants in Germany.

5.3 Results

The results are split into two sections. First, a comparison between the KF and the SKF was carried out at each site to demonstrate the SKF approach for correcting the bias present between the CAMS and TOAR datasets. Next, the performance of the SKF was evaluated for infilling missing data. Three types of missing data are considered: random missingness, consecutive missingness and a real world scenario, which provides a combination of random and consecutive missingness.

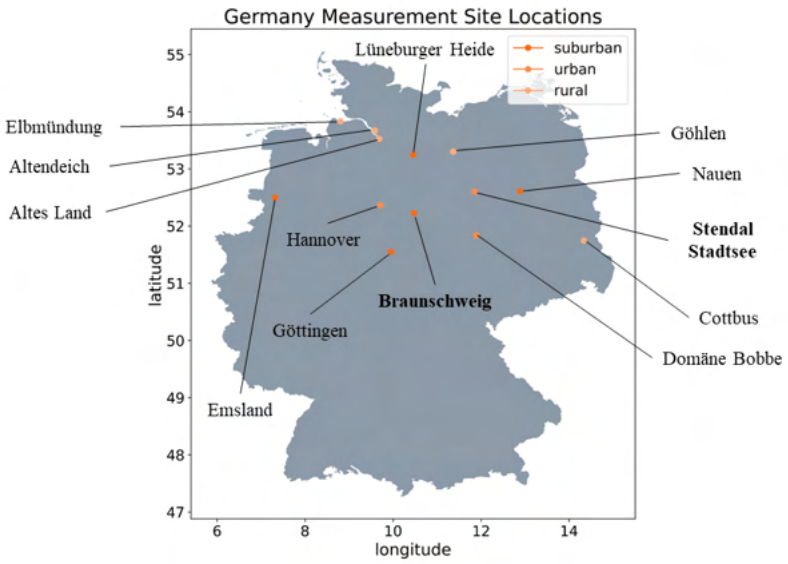


Figure 5.2.2: Map of measurement site locations in Germany, colour indicates site type. Sites in bold are on both the Airbase and UBA networks, remaining sites are on the Airbase network only.

5.3.1 Correcting the Bias

Figure 5.2.1b shows the difference between the CAMS and TOAR concentrations for a suburban background site located in Broitzem, Braunschweig 2015 to 2017. The average CAMS bias is highest in the summer months, June to September, and while CAMS is able to capture ozone levels reasonably well in the winter months, December to February, it is still biased. Similar trends are seen across all sites, with the site specific RMSE between CAMS and TOAR varying between 5.1 ppb and 5.9 pbb.

The CAMS data, which we treat as noisy observations of the TOAR ozone data, are not symmetric around the TOAR data due to the bias present, which means that the KF is fundamentally unable to capture the underlying process as it requires the observations to be normally distributed. Figure 5.3.1a shows the results for the KF

when applied to the CAMS dataset. The RMSEs between CAMS and TOAR and between the KF and TOAR are 5.1 ppb and 5.2 ppb, respectively. Therefore, there is no benefit in using a KF as it does not reduce the bias. Figure 5.3.1b shows the fit of the SKF, which results in a marginal improvement in RMSE, 5.0 ppb compared to CAMS, 5.1 ppb. This small improvement is due to the temporal variability in the skewness.

As the accuracy of CAMS varies throughout the year, a sliding window approach was used to vary the skew over time. After considering window lengths, 10 days was chosen as it gave the greatest reduction in error when compared to the observation data. The noise parameters, W and V, are estimated across the whole dataset, the skew parameter, λ, is then estimated over a 10 day window starting from 1st January 2015, followed by a 10 day window starting on 2nd January 2015 etc. The noise parameters are then updated accordingly, and this is repeated for the length of the dataset resulting in a skew estimate for each day. Figure 5.3.1c shows the SKF applied to the CAMS data with varying skew. This results in a SKF estimate that corrects the CAMS concentrations to a level more consistent with the TOAR data, with the RMSE between the variable skew Kalman filter and TOAR being 4.6 ppb.

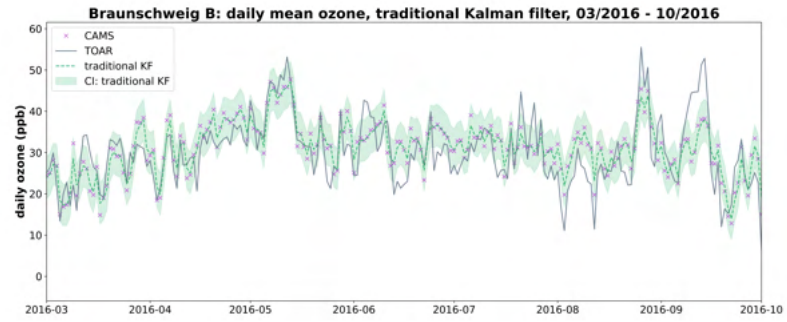
Both SKFs perform better than the raw CAMS data and the KF. As expected, the varying SKF performs best overall. Figure 5.3.1d shows the results for all filters from the beginning of March 2016 to end of September 2016. From this we can see the KF often sits too high to capture the TOAR concentrations. This is because it is fitting to the mean of the CAMS data and so cannot correct for any bias present. The reduction in RMSE between the raw CAMS and the constant SKF is 0.1 ppb. This is due to an invalid assumption that the skew is constant, resulting in a low estimated skew across the entire dataset. This is further shown in the estimated skew in Figure 5.3.1d for the sliding window approach: while the skew is often close

Table 5.3.1: Comparison of RMSEs for CAMS and two filters for all sites. Lowest values for each method in bold.

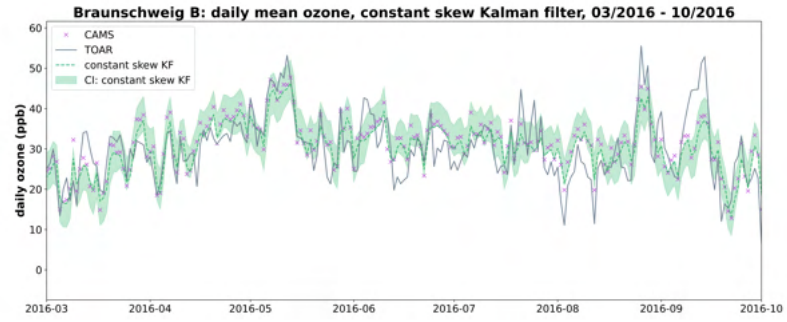
Station Name	**RMSE raw CAMS (ppb)**	**RMSE Constant SKF (ppb)**	**RMSE Variable SKF (ppb)**
Emsland	5.2	5.4	**5.0**
Stendal Stadtsee B	5.4	5.4	**4.8**
Göttingen	5.7	5.5	**5.2**
Domäne Bobbe	5.7	5.8	**5.0**
Cottbus	5.4	5.4	**5.0**
Lüneburger Heide	5.8	5.7	**5.1**
Altes Land	5.9	5.2	**4.4**
Elbmündung	5.5	5.4	**4.7**
Göhlen	5.7	5.5	**4.7**
Altendeich	5.9	5.4	**4.6**
Stendal Stadtsee A	5.5	5.5	**4.9**
Braunschweig B	5.1	5.1	**4.6**
Hannover	5.7	5.3	**4.9**
Braunschweig A	5.2	5.1	**4.7**
Nauen	5.8	5.2	**4.5**

to zero, it has periods of high and low skew that cannot be captured by a constant skew parameter. The reduction in RMSE between the raw CAMS and the varying skew Kalman filter is 0.5 ppb. Visually this improvement can be seen in Figure 5.3.1d, where the variable SKF is often sitting much lower that the constant SKF in sections where there is a greater bias, notably at the beginning of May and August.

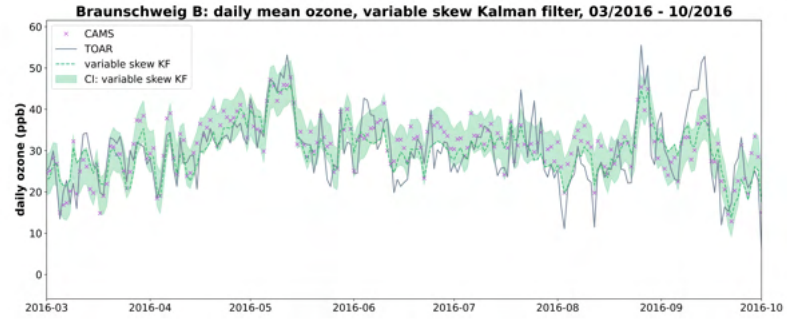
Lastly, Table 5.3.1 shows a comparison of the RMSE error for the raw CAMS,



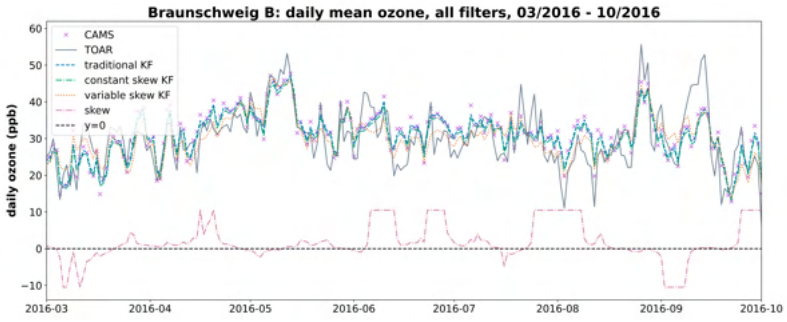
(a)



(b)



(c)



(d)

Figure 5.3.1: Output of filter estimates **(a)** traditional Kalman filter, **(b)** skew Kalman filter where the skew is assumed to be constant in time, **(c)** skew Kalman filter where the skew is calculated over a 10 day sliding window, **(d)** comparison of the three filters and skew varying in time. The 95% confidence interval for each filter is given by the shaded area.

constant and variable skew Kalman filter for all 15 sites in Germany. The varying skew filter performs the best, reducing the error with TOAR at every site. The mean improvement to RMSE between the raw CAMS and the variable skew filter estimate is 0.76 ppb.

5.3.2 Infilling Missing Data

5.3.2.1 Randomly Missing Data

Next we demonstrate an infilling method for missing data obtained from the corrected observations from CAMS with the varying SKF model. Starting from a complete dataset, using the same site and time period as before, we removed varying percentages of the data at random. Any days with missing data were ignored and the model was fit to the remaining data using the varying skew approach outlined above, giving a skew estimate at each time step of available TOAR data. The day of year mean is then used to produce a complete set of skew estimates. The bias between the datasets varies seasonally, as shown in Figure 5.2.1b, meaning that, the skew also follows seasonal trend. In the event that the same day is missing from every year we simply linear interpolate the data to ensure we have a skew estimate for every day.

Four scenarios are considered with either 5%, 10%, 15% or 20% of the data missing at random, with the SKF estimate compared to both the CAMS raw data and linear interpolation. Performance is evaluated using 25 or 100 runs of randomly missing data at each percentage, with the difference between the RMSE of the skew Kalman method and the RMSE of each of the raw CAMS or linear interpolation used to determine which method performs best.

Figure 5.3.2 shows the infilling results for 15% missing data for a 2 month period at the Braunsweig site. An immediate advantage to using the SKF approach as

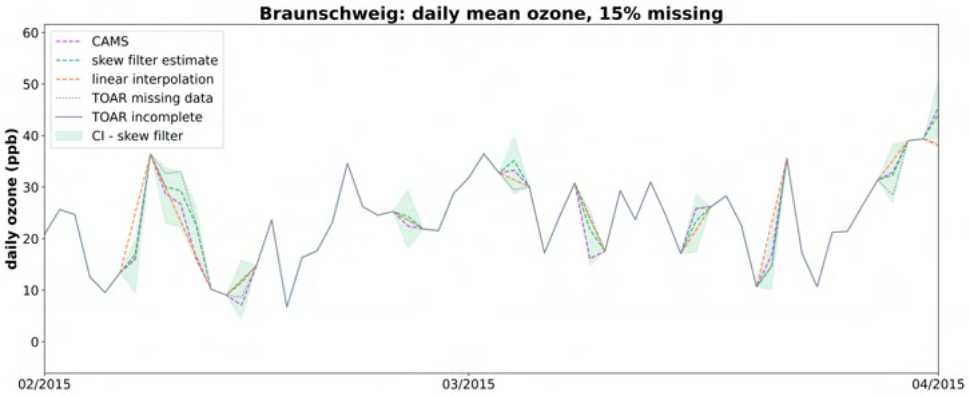


Figure 5.3.2: Comparing infilling methods (dashed lines) for 15% randomly missing data. The TOAR data (grey) is shown with the known missing data plotted as dotted line. CAMS purple, linear interpolation orange, skew filter estimate green, and the 95% confidence interval for the skew filter is shown as the shaded green area.

opposed to linear interpolation, is that we have an associated uncertainty with the infilled values. Further linear interpolation fails to match the structure of the data, which will result in increasing errors longer periods of missingness; both CAMS and the SKF estimate are able to better capture this.

Figure 5.3.3 shows histograms of the difference in RMSE between each of linear interpolation and CAMS and the filter estimate for each run of missing data for the Braunsweig site. Positive values indicate when the skew filter estimate is performing better than the compared method. We also consider the mean difference in error, which is the difference in the mean RMSE over the runs of the compared method with the SKF. Here, a positive value again implies the SKF is performing better and indicates the magnitude of the improvement. With 5% missingness the SKF estimate performs better in 66% and 79% of cases compared to linear interpolation and CAMS respectively, with a mean difference of 0.2 ppb and 0.3 ppb. With 20% missing data, the SKF performs best in 77% and 65% of cases compared to linear interpolation and CAMS respectively, with a mean difference of 0.3 ppb and 0.1 ppb. Thus, while the skew filter method does not always perform better, it

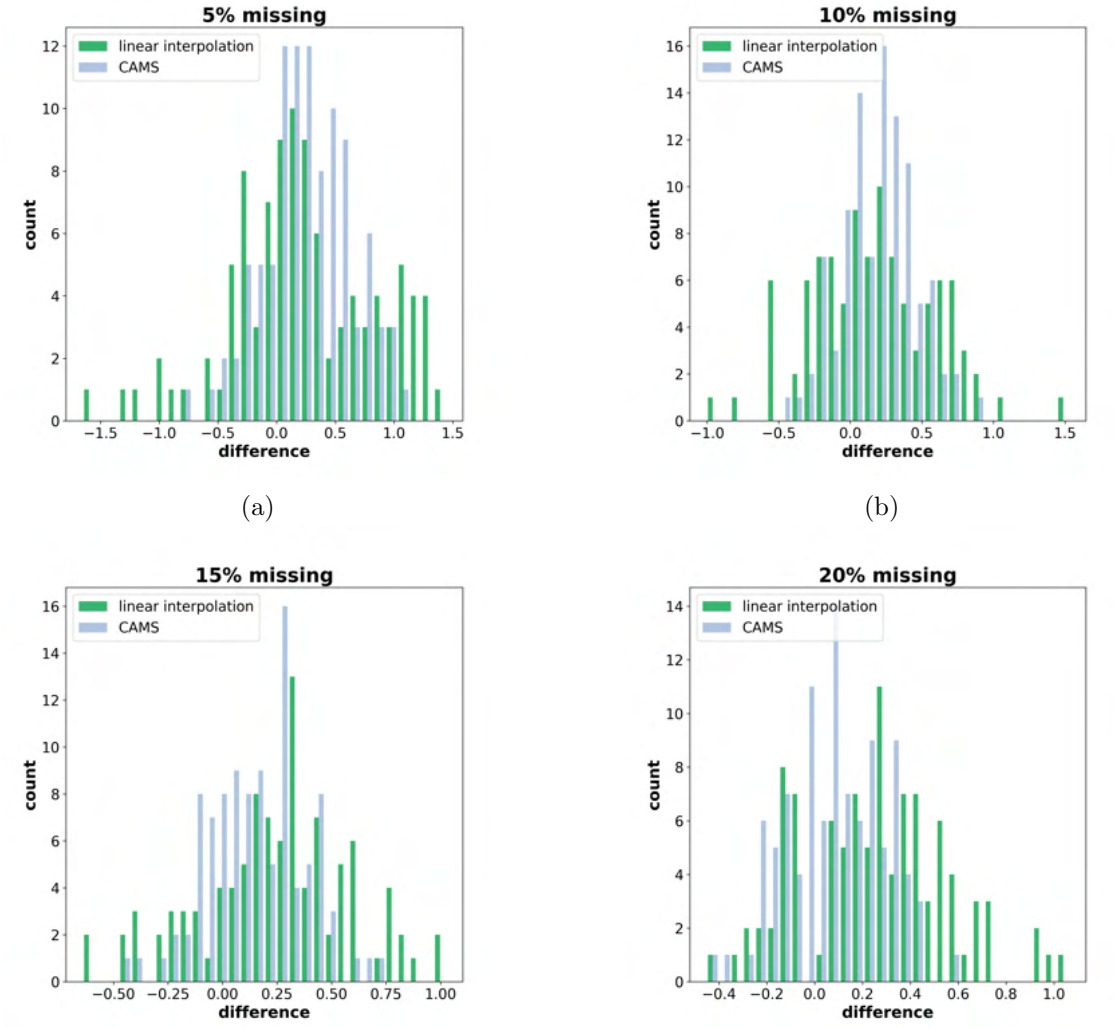


Figure 5.3.3: Difference in RMSE for **(a)** 5%, **(b)** 10%, **(c)** 15% and **(d)** 20% missing data. The difference between linear interpolation and the skew filter estimate is shown in green, the difference between CAMS and the skew filter estimate is shown in grey. Positive values indicated the skew filter estimate has a lower RMSE when compared to the known missing TOAR data.

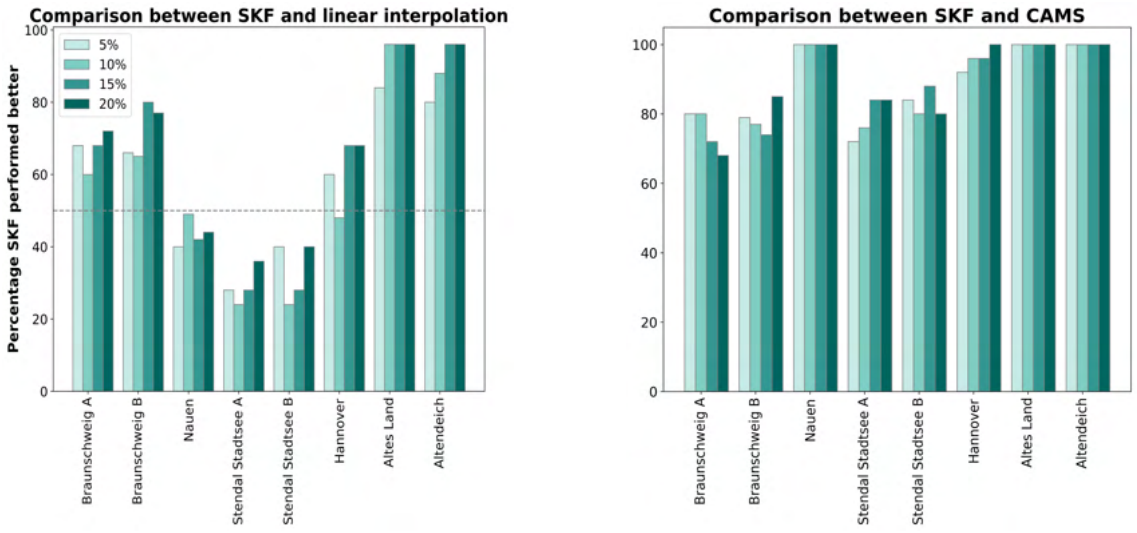
performs better more often than not. The differences between the errors show that when the SKF method is performing better it is often by a non-negligible amount, i.e. there are tangible gains compared to using the CAMS data directly or by linear interpolation of the TOAR observations.

Extending our analysis to more sites, Figures 5.3.4a and 5.3.4b show the percentage of times the filter estimate had a lower error than each of the comparison method. The SKF approach consistently performs better than the raw CAMS data, with a mean difference of between 0.1 ppb and 1.2 ppb across sites. Performance compared to linear interpolation was more mixed, with the mean difference between -0.5 ppb and 0.6 ppb, although the SKF approach performed better the majority of the time at 5 of the 8 sites. Figure 5.3.5 compares performance of each method for the varying amounts of missing data. For the majority of sites, as the amount of missing data is increased the SKF approach shows an increasing improvement over linear interpolation, whereas performance remains the same or starts to decrease compared to CAMS.

5.3.2.2 Consecutive Missing Days

In practice, consecutive missing days are more often seen in air quality datasets than single missing values. For Ashton Hill, a UK based site, 31% of the missing observations are single incidences of missing data over the same time period. The performance for 3, 5 and 7 days of consecutive missing days throughout the dataset is evaluated. As before, we consider 25 runs using data over the same time period as before, although this time where 5 incidences of consecutive missing days are removed at random. Incidences are constructed so as not to overlap. As before, the difference in error between the SKF estimate and both CAMS and linear interpolation are considered.

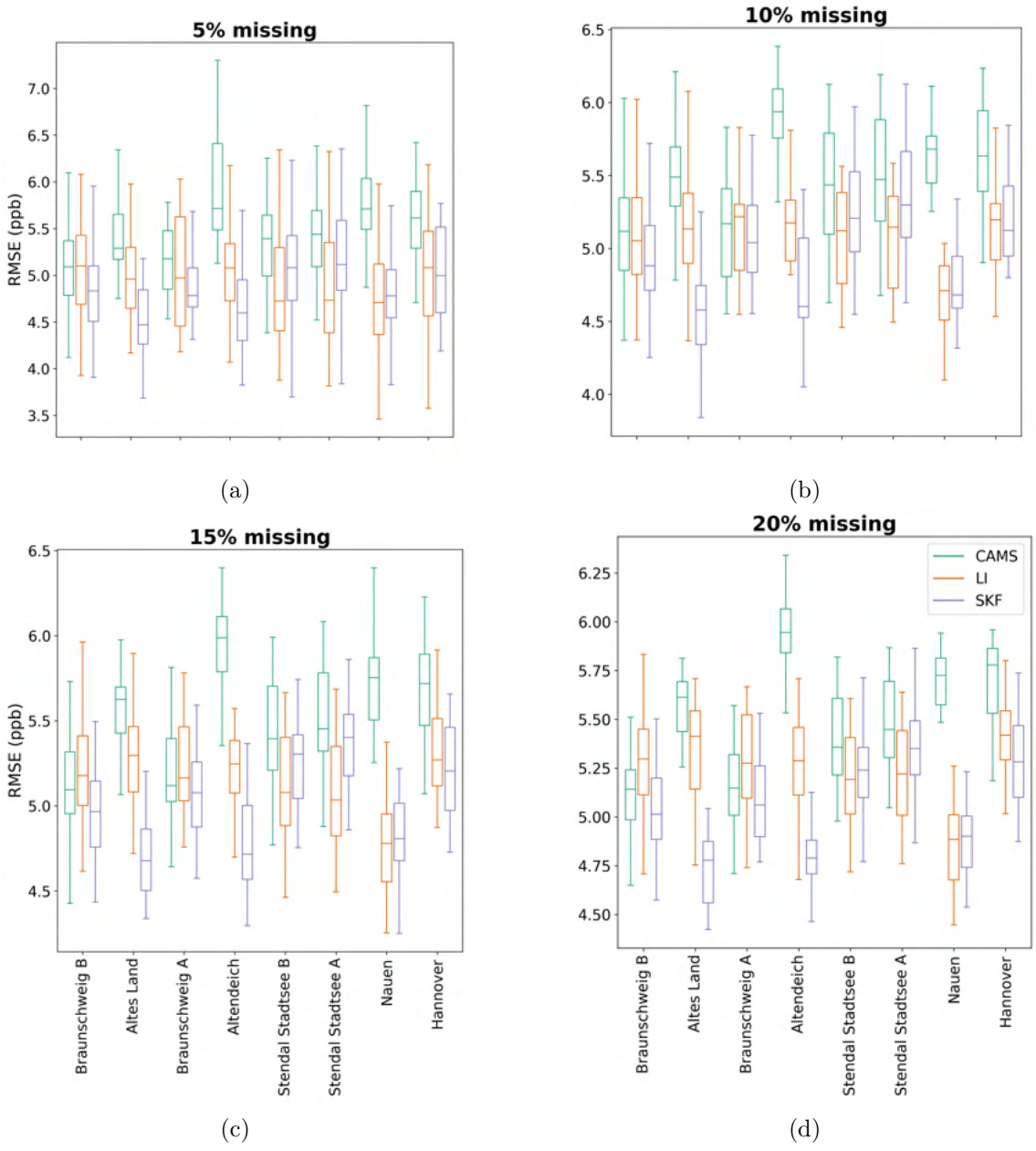
Using the same sites, Figure 5.3.6 summarises the results across the three periods



(a) (b)

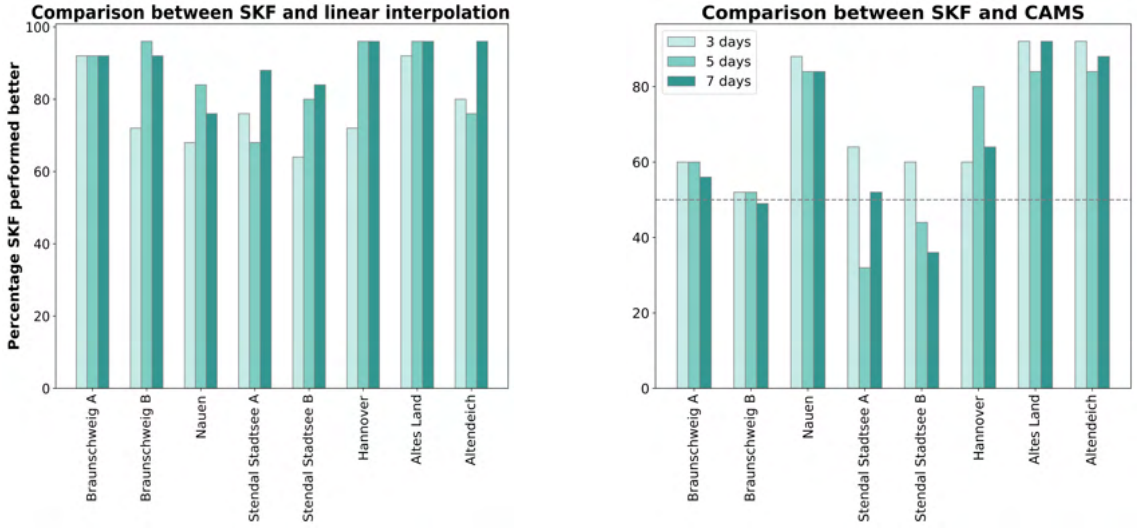
Figure 5.3.4: Percentage the skew Kalman filter performed better than **(a)** linear interpolation or **(b)** for randomly missing data of 5%, 10%, 15% and 20%. The dashed grey line indicates 50%, where sites falling below this line the skew Kalman filter performs better less than half the time.

of missing data. The SKF method consistently performs better than linear interpolation at each site with a mean difference in error between 0.8 ppb and 1.8 ppb for 3 day periods, 1.1 ppb and 2.4 ppb for 5 day periods, and 1.6 ppb and 2.6 ppb for 7 day periods. Compared to the raw CAMS data, the SKF performs better at 5 of the 8 sites across the increasing periods of missing data. For 3 days consecutive missing the SKF performs best at all sites, with a mean difference in error between 0.2 ppb and 1.0 ppb. The mean difference in error is between -0.2 ppb and 1.4 ppb for 5 day periods and -0.1 ppb and 1.2 ppb for 7 day periods. Figure 5.3.7 compares the performance of each method for the varying amounts of missing data. When compared to linear interpolation, the SKF performs increasingly better or stays the same as the length of consecutive missing days is increased. Conversely, the SKF performs worse or the same as the length of consecutive missing days is increased compared to the raw CAMS data.



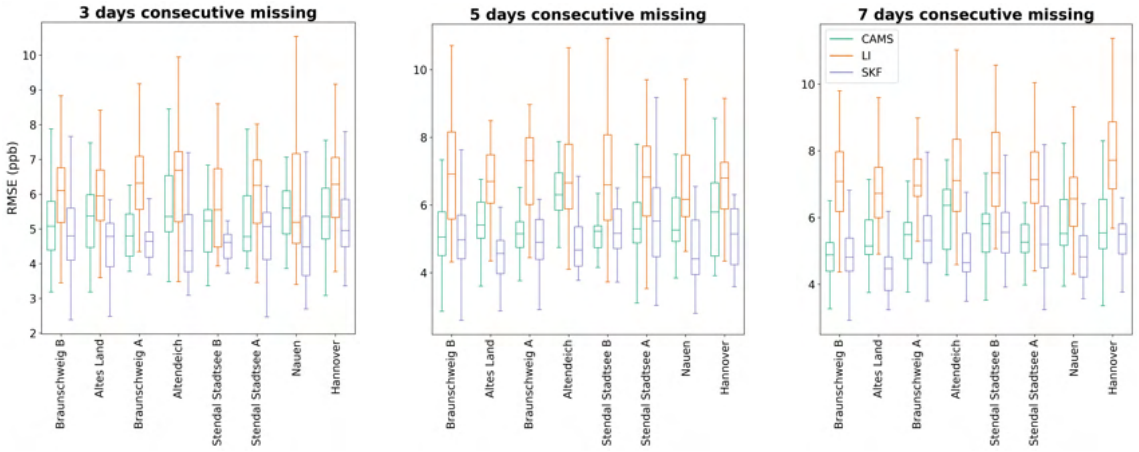
(a) (b) (c) (d)

Figure 5.3.5: Performance of CAMS (green), linear interpolation (LI) (orange) and the skew Kalman filter (purple) for infilling randomly missing data of **(a)** 5%, **(b)** 10%, **(c)** 15% and **(d)** 20%.



(a) (b)

Figure 5.3.6: Percentage the skew Kalman filter performed better than **(a)** linear interpolation or **(b)** for consecutive missing data of 3, 5, and 7 days. The dashed grey line indicates 50%, where sites falling below this line the skew Kalman filter performs better less than half the time.



(a) (b) (c)

Figure 5.3.7: Performance of CAMS (green), linear interpolation (LI) (orange) and the skew Kalman filter (purple) for infilling consecutively missing data of **(a)** 3, **(b)** 5, and **(c)** 7 days.

Table 5.3.2: Results for real world scenario.

Site	**RMSE raw CAMS (ppb)**	**RMSE Linear Interpolation (ppb)**	**RMSE Skew Kalman Filter (ppb)**
Braunschweig A	6.2	6.8	**5.7**
Braunschweig B	6.2	6.7	**5.7**
Nauen	5.8	6.6	**4.9**
Stendal Stadtsee A	5.8	7.6	**5.6**
Stendal Stadtsee B	5.7	7.4	**5.6**
Hannover	6.1	6.7	**5.3**
Altes Land	5.7	6.1	**4.4**
Altendeich	6.1	6.4	**4.9**

5.3.2.3 Real World Scenario

Typically, datasets with missing data contain both singular missing values as well as periods of consecutively missing data. To demonstrate our approach for real world missing data we take the occurrences of missing data from another site over the same time period and overlay that over the 8 sites considered before. Aston Hill is a rural background site on the Automatic Urban and Rural network (AURN; https://uk-air.defra.gov.uk/data/), which is an hourly air quality database in the UK. This site was chosen because it has considerable missing data, yet does not exceed more than 7 consecutive days of missing data and has at least 80% data coverage. Between 2015-2017, 5.7% of the data is missing, with 29 incidences of missing data, of which 9 are single missing values. The longest gap is 5 days and there are two occurrences at this length.

The incidences of missing data are generated in each of the 8 sites used in the previous sections and performance is evaluated using RMSE. The results are given in Table 5.3.2. The SKF approach performs best at every site, with the largest improvement seen compared to linear interpolation with a difference in error between 1.0 ppb and 2.0 ppb. Compared to the raw CAMS data there is a reduction in error between 0.1 ppb and 1.3 ppb.

5.4 Discussion and Conclusion

We have proposed a method for infilling missing measurement data using bias corrected reanalysis data, conceptualising the bias as a skew facilitating implementation of a SKF. We have demonstrated our method using surface ozone data. We explored the performance of the method at 8 different measurement sites. The performance varied across measurement sites types, which warrants further exploration and potentially using covariates.

Our method consistently performed better in comparison to linear interpolation for consecutive missing data and when compared with the raw CAMS data, performed better for randomly missing data. While compared to linear interpolation it does not always perform better for random missingness it does have the added benefit of quantifying the uncertainty in the estimate which linear interpolation lacks. The performance compared to CAMS decreased as the amount of missing data increased. For the real world scenario, which offers a realistic combination of single and consecutive missing values, our skew Kalman approach performed better at every site. The greatest improvement was seen when compared to linear interpolation. The lowest reduction in error was at sites which the skew Kalman method did not perform as well for consecutive missing data.

As this method has not been tailored to the air quality problem it is applicable across other applications in which a secondary bias dataset is available. The simplicity and computational efficiency of this method is the main benefit of this approach, as well as the transferability to other applications. One drawback to the varying skew approach is that it increases the computational time of the model as a result of calculating the skew estimate at each time step. However, it is still relatively quick, and fits on the order of minutes.

While our current implementation prioritises the simplicity of using the two datasets without additional information, there is scope to further improve the method with the inclusion of covariates, in the case of ozone, useful covariates include temperature and other meteorological variables (Otero et al. 2016). Another natural extension would be to make use of the information around the gap in the observation data. Again, to preserve the simplicity of the skew Kalman filter approach we only used the reanalysis data when infilling the missing data.

Thus, when selecting methods for infilling missing data, suitability of the application

should be considered. For large amounts of missing data, or long periods of consecutive missing data, this method may not be the best. However, for data with reasonable coverage and short periods of missing data, this method can perform well.

Chapter 6

A Comparison of Approximate Methods for Big Data Spatial Inference

6.1 Introduction

Large datasets can be problematic for modern statistical applications using standard computational techniques for Bayesian inference as estimating the model parameters requires repeated evaluations of the likelihood function which is computationally expensive. Often large datasets are too big to process on a single machine due to the processor, memory or disk limits. Processor bottlenecks can be handled by a graphics processing unit, however, memory or disk limits can only be alleviated by splitting data across multiple machines. Thus, to carry out Bayesian inferences on large datasets it is practical to divide the data into subsets and carry out the inference on the subsets. Thus, we require methods in which we can combine the output from these subsets. Communication between machines is expensive regardless of how much data is being communicated, thus algorithms must be able to perform distributed Bayesian analyses with minimal communication.

Gaussian processes have been the main tool used for analysing geospatial data. The appealing properties of the Gaussian distribution have led to Gaussian processes becoming an indispensable tool for any spatial data analyst to perform tasks such as spatial prediction and proper uncertainty quantification. Gaussian processes struggle with computation intractability for large data sets. Evaluating the density requires $O(N^3)$ operations which can become challenging for moderately large N, where N is the number of data points. An alternative is to build a low rank approximation to the covariance matrix based around 'inducing variable' (Quiñonero et al. 2005, Titsias 2009). For these low rank approximations, the computational complexity is now $O(nm^2)$ where m is the number of inducing variables selected by the user (Hensman et al. 2013).

Recent work has focussed on the efficient use of modern computational platforms and the development of methods that are parallelize-able. These include using parallel computing to calculate the density function for spatial Gaussian processes, the use of a basis function approach that lends itself to distributed computing and using only nearest neighbours to factorise the density function as a series of conditional distributions. Another approach looks at dividing the data in a large number of subsets, doing inference on the subsets in parallel and then combining the inference, Guhaniyogi et al. (2019) present a divide and conquer approach called distributed Krigging for GP-based spatial models to alleviate the computational challenges of large scale spatial inference.

Various methods for combining the subsets through a divide and conquer approach exist in the independent setting, including consensus Monte Carlo (Scott et al. 2016), Gaussian Process barycentres (Mallasto & Feragen 2017), and SwISS (Sub-posteriors with Inflation, Scaling and Shifting) (Vyner et al. 2022). However, as these make assumptions around independence between the data subsets which do not hold in the spatial setting, the combined inference will not be exact. A necessary first step in

establishing a method for large-scale spatial inference is evaluating the performance of these methods in the spatial setting. In this work, we consider a GP-based model and evaluate the performance of the combining methods consensus Monte Carlo, Gaussian Process barycentres, and SwISS for approximate inference carried out on the full data set. While this is not the intended use for these algorithms, and the independence assumption is not satisfied, we demonstrate that these methods offer a reasonable approximation for carrying out parameter inference on the full dataset.

Initially, we evaluate the methods on a small to moderate-sized simulated data set with 625 spatially referenced data points. We use a Gaussian Process to model the data, and the kernel length scale and variance are estimated using Markov Chain Monte Carlo (MCMC). Performance is evaluated by considering the runtime of each method and how well it captures the full data fit. We also consider the predictive accuracy of the methods by splitting the data into training and test portions and evaluating how well it captures the unseen data. For our divide and conquer approach the training portion of the data is then divided into subsets. We then demonstrate our method on a larger real-world data set. Here we use average USA temperature data as this offers a reasonably large dataset, 5660 spatially referenced data points, where it is still feasible to fit the full dataset but implementing a divide and conquer approach significantly alleviates the computational time.

This chapter is structured as follows, Section 6.2 introduces the model and briefly describes each of the combining methods, Section 6.3 evaluates the perforce of the methods for simulated data and a real world example. Finally, in Section 6.4 we discuss our results and present our conclusions.

6.2 Methods

6.2.1 Gaussian Process Regression

Gaussian process regression (GPR) (Rasmussen & Williams 2006) is a Bayesian approach to regression that works well for small datasets and also provides uncertainty measurements on the predictions. Rather than learn exact values for each parameter we infer a probability distribution over all possibly values. We are interested in modelling the relationship between N observations $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ at corresponding inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ using a model f with which we can make predictions at an unseen set of test points X^*. $f(x) \sim GP(\mu, k)$ means that for any finite collection of functions values $\mathbf{f} = f(\mathbf{X})$

$$\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2, ..., f(\mathbf{x}_n)] \sim N(\mu, K) \tag{6.2.1}$$

To make predictions at unobserved test points based on observed training points we assume the following

$$\begin{bmatrix} f \\ \\ f* \end{bmatrix} \sim N\left(\mu, \begin{bmatrix} K_{f,f} + \sigma^2 I_n & K_{f^*,f} \\ \\ K_{f,f^*} & K_{f^*,f^*} \end{bmatrix}\right), \tag{6.2.2}$$

where I_n denotes an $n \times n$ identity matrix. We denote function values at the test points as $f^* = f(X^*)$. Using the properties of the normal distribution, the conditional distribution of the test data is

$$\begin{aligned} Cov(f^*) &= K_{f^*,f^*} - K_{f,f^*}(K_{f,f} + \sigma^2 I_n)^{-1} K_{f^*,f}, \\ \mu(f^*) &= K_{f,f^*}(K_{f,f} + \sigma^2 I_n)^{-1} f. \end{aligned} \tag{6.2.3}$$

From these we can obtain the predictive mean and covariances at the test points conditioned on the training data. To use the equations above we must model the

covariances and cross-covariances between training and test data. To do this we must specify a GP with a mean function, $m(x_i)$, and covariance function, $k(x_i, x_j)$,

$$f(x_i) \sim GP(m(x_i), k(x_i, x_j)). \tag{6.2.4}$$

Within this GP prior, we can incorporate prior knowledge about the space of functions through the selection of the mean and covariance functions. The covariance function $k(x, x')$, models the dependence between the function values at input points x_i and x_j. The function k is commonly called the kernel. The appropriate choice of kernel is based on assumptions such as smoothness and likely patterns expected in the data. A sensible assumption is that the correlation between data points decays with the distance between the points, thus a common choice on kernel is the radial basis function kernel, which is defined as

$$k(x_i, x_j) = \sigma_f^2 \exp\left(- \frac{||x_i - x_j||^2}{2\lambda^2} \right). \tag{6.2.5}$$

The radial basis function provides an expressive kernel to model smooth and stationary functions. The two hyper-parameters, length-scale λ, and signal variance σ_f^2 can be varied to increase or reduce the a priori correlation between points and consequently the variability of the resulting function. In practice, these parameters are estimated using Markov Chain Monte Carlo or MLE and optimising the likelihood function. This is carried out by minimising the negative log likelihood of the hyper parameters $\theta = (\sigma_f, \lambda)$

$$\begin{aligned} p(\theta|y) &= \int p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}|X) d\mathbf{f} \\ &= -\frac{1}{2}\mathbf{y}'(K_\theta + \sigma_n^2 I)^{-1} - \frac{1}{2}\log|K_\theta + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \end{aligned} \tag{6.2.6}$$

Once the parameters are optimised this can be used in the predictive equations for the mean and covariance to obtain estimates at the test points. Since both the prior and likelihood are Gaussian, the integral can be evaluated in the closed form to give

the Gaussian predictive distribution,

$$p(f^*|\mathbf{y}) = N(K_{f^*,f}(K_{f,f} + \sigma_\epsilon I)^{-1}\mathbf{y}, K_{f^*,f^*} - K_{f^*,f}(K_{f,f} + \sigma_\epsilon^2 I)^{-1}K_{f,f^*}). \quad (6.2.7)$$

Calculating $p(f^*|\mathbf{y})$ requires the inversion of an $n \times n$ matrix which requires $O(n^3)$ operations, where n is the number of training data points. Due to the high computational cost involved in inverting this matrix an exact implementation can only handle problems at most of a few thousand training cases practically on today's machines (Quiñonero-Candela et al. 2007). Our GPR is carried out using the *GPJax* package (Pinder & Dodd 2022).

6.2.2 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms for sampling from a posterior distribution. To achieve this, we start from some initial point and draw samples from a proposal distribution which is some easy to simulate distribution. We determine whether this sample is from the posterior using the Metropolis-Hastings acceptance probability. These steps are repeated forming a Markov chain where the next sample depends on the previous. Samples are drawn until the Markov chain forms a stationary distribution which is the posterior. The earlier samples are discarded, often referred to as 'burn in'. We implement the No U-turn sampler Hoffman & Gelman (2014) using the *Blackjax* package (Lao & Louf 2020).

6.2.3 Methods of Combining

Let $f(\mathbf{y}|\theta)$ be the likelihood for a statistical model, parametrised by $\theta \in \mathbb{R}^d$, for a data set $\mathbf{y} = \{y_1, y_2, ..., y_n\}$ of length n. Let $p_0(\theta)$ denote the prior density for the parameter vector θ, then our posterior density is, up to a constant of proportionality,

$$p(\theta|\mathbf{y}) \propto p_0(\theta)p(\mathbf{y}|\theta). \quad (6.2.8)$$

We assume that $\mathbf{y}$ is independent and can be partitioned into S subsets, $\mathbf{y}_1, ..., \mathbf{y}_S$, such that the likelihood for the full data is the product of the likelihoods for the individual subsets. The posterior density for θ given $\mathbf{y}$ is, up to a constant of proportionality

$$p(\theta|\mathbf{y}) \propto p_0(\theta) \prod_{s=1}^{S} p_s(\mathbf{y}_s|\theta). \tag{6.2.9}$$

We will focus on three methods of combining, Gaussian process barycentres, Consensus Monte Carlo, and SwISS. Here, we will give a brief introduction to each method.

6.2.3.1 Consensus Monte Carlo

The Consensus Monte Carlo Algorithm (Scott et al. 2016) describes a method of performing approximate Monte Carlo simulation from a Bayesian posterior distribution based on very large data sets.

The data is split into groups called 'subsets', each subset is given to a worker machine which does a full Monte Carlo simulation from a posterior distribution given its own data, and then the posterior simulations from each worker are combined to produce a set of global draws representing the consensus belief among all the workers.

Let y represent the full data and y_s denote a subset, and let θ denote the model parameters. For a model with the appropriate independence structure the posterior density can be written

$$p(\theta|\mathbf{y}) \propto \prod_{s=1}^{S} p(\mathbf{y}_s|\theta) p(\theta)^{1/S} \tag{6.2.10}$$

The prior is broken into S components to preserve the total amount of prior information in the posterior. It is assumed that the batches of observations are independent across the subsets, given the parameters, but it allows for arbitrary

dependence within the elements of $\mathbf{y}_s$.

If each worker s generates G draws $\theta_{s1}, ..., \theta_{sG}$ from $p_s(\theta|\mathbf{y}) \propto p(\mathbf{y}_s|\theta)p(\theta)^{1/S}$, we can combine draws using weighted averages. If each worker is assigned a weight represented as a matrix W_s the consensus posterior for draw g is $\theta_g = (Var(\theta|\mathbf{y}_s)W_s)^{-1}(Var(\theta|\mathbf{y}_s)W_s\theta_{sg})$.

The weight $W_s = \sum_s^{-1}$ is optimal for Gaussian models.

Algorithm:

1. Divide $\mathbf{y}$ into subsets $\mathbf{y}_1, ..., \mathbf{y}_s$.
2. Run S separate Monte Carlo algorithms to sample $\theta_{sg} \sim p(\theta|\mathbf{y}_s)$ for $g = 1, ..., G$, with each subset using the fractioned prior $p_0(\theta)^{1/S}$.
3. Combine the draws across the subset using weighted averages: $\theta_g = (Var(\theta|\mathbf{y}_s)W_s)^{-1}(Var(\theta|\mathbf{y}_s)W_s\theta_{sg})$.

The algorithm is exact only for Gaussian posteriors. However, it does work well when applied to non-Gaussian posteriors (Scott et al. 2016).

6.2.3.2 SwISS

The SwISS algorithm (Vyner et al. 2022) is an alternative to the consensus Monte Carlo algorithm which recombines the data differently but is still computationally quick to run, exact in the Gaussian case, does not require tuning and scales well to higher dimensions.

Compared to the consensus approach, SwISS does not merge samples, but instead applies a transformation to the posterior samples that are generated from a stochastic approximation of the full posterior. The stochastic approximation is

referred to as the inflated sub-posterior, which is a posterior density, conditional on a subset of the data, raised to a positive power.

$$p_s(\theta|\mathbf{y}_s) \propto p_0(\theta)p(\mathbf{y}_s|\theta)^S \tag{6.2.11}$$

Inflating the sub-posterior in this manner has the effect of approximately preserving the shape of the posterior density conditional on the full data. Affine transformations (shift and re-scale) are applied to the algorithm of (Wu & Robert 2017) which shifts each sub-posterior.

1. Divide $\mathbf{y}$ into subsets $\mathbf{y}_1, ..., \mathbf{y}_s$.
2. Run S separate Monte Carlo algorithms to get J samples $\{\theta_s^j\}_{j=1}^J$ from each of the S inflated posteriors $p_s(\theta|\mathbf{y}_s) \propto p_0(\theta)p(\mathbf{y}_s|\theta)^S$
3. Calculate the mean μ_s and variance V_s for each inflated sub posterior
4. Set the global mean μ and variance V and calculate the matrix square root

$$\mu = V\frac{1}{S}\sum_{s=1}^{S} V_s^{-1}\mu_s \qquad V = (\frac{1}{S}\sum_{s=1}^{S} V_s^{-1})^{-1} \qquad M = SPSQ(V) \tag{6.2.12}$$

   where $SPSQ(V)$ denotes the symmetric positive-definite square root of the matrix V.

5. For each $s \in \{1, ..., S\}$ apply the affine transformations to the inflated posterior samples
   - $\tilde{V}_s = M^{-1}V_sM^{-1}$
   - $\tilde{M}_s = SPSQ(\tilde{V}_s)$
   - $A_s = M\tilde{M}_s^{-1}M^{-1}$
   - $\theta_s^{1:J} = A_s(\theta_s - \mu_s) + \mu$
6. Concatenate the transformed samples $\theta_s^{1:J}$ to give a Monte Carlo approximation of the full posterior distribution $p(\theta|\mathbf{y})$

6.2.3.3 Barycentres

The 2-Wasserstein distance metric between two probability measures μ and ν is defined as the optimal cost required to transport the unit mass from μ to ν, or vice-versa. When μ and ν are both multivariate Gaussian distributions the solution is analytically given by

$$W_2^2(\mu,\nu) = ||m_1 - m_2||_2^2 + Tr(V_1 + V_2 - 2(V_1^{1/2}V_2V_1^{1/2})^{1/2}) \tag{6.2.13}$$

where $\mu \sim N(m_1, V_1)$ and $\nu \sim N(m_2, V_2)$

For a collection of T measure $\{\mu_i\}_{t=1}^T \in P_2(\theta)$, the Warrestein barycentre $\bar{\mu}$ is the measure that minimises the average Wasserstein distance to all other measures in the set. We can write this as the Fréchet mean on a Wasserstein space

$$\bar{\mu} = argmin_{\mu \in P_2(\theta)} \sum_{t=1}^{T} \alpha_t W_2^2(\mu, \mu_t), \tag{6.2.14}$$

where α_t is a weight vector that sums to 1. The Wasserstein barycentre, $\bar{\mu}$, is often a computationally demanding optimisation problem. As with consensus Monte Carlo, when the measures are multivariate Gaussian, the barycentre $\bar{\mu} = N(\bar{m}, \bar{V})$ has analytical solution

$$\begin{aligned} \bar{m} &= \sum_{t-1}^{T} \alpha_t m_t, \\ \bar{V} &= \sum_{t-1}^{T} \alpha_t (\bar{V}^{1/2} V_t \bar{V}^{1/2})^{1/2}. \end{aligned} \tag{6.2.15}$$

We can identify $\bar{V}$ using a fixed-point iterative update.

Mallasto & Feragen (2017) show that the barycentre $\hat{f}$ of a collection of Gaussian processes $\{f_i\}$ such that $f_i \sim GP(\bar{m}_i, \bar{V}_i)$ is non degenerate for any finite set of GPs

$\{f_t\}_{t=1}^{T}$ i.e., $T < \infty$. Thus, for an n-dimensional finite Gaussian distribution $f_{i,n}$, the Wasserstein metric between any two Gaussian distributions $f_{i,n}, f_{j,n}$ converges to the Wasserstein metric between GPs as $n \to \infty$. Assuming a Gaussian approximation for each inflated sub-posterior, the barycentre is the geometric centre of the inflated sub-posterior distributions. The R library waspr (Cremers 2020) was used to combine the subsets.

6.3 Results

6.3.1 Simulation Study

Before implementing the divide and conquer approaches with real data, we evaluate the effectiveness of each combining method using simulated data. The data, shown in Figure 6.3.1, is simulated from a GP using a Matern kernel over a grid with 625 points. We use x and y for our coordinate space to simulate the observations, both are fixed between -5 and 5. While 625 points is well within the limit of data points that can be fit on a single machine it allows us to easily compare the output from each of the combining methods to the full data fit. For each of our approximate methods we divide the data into subsets. We will consider 2 cases, splitting the data into 2 subsets of 312 and 313 data points and splitting the data into 5 equal subsets of 125 data points. Each subset is a randomly sampled, non-overlapping subset of the full dataset.

Figure 6.3.2 shows the expected value for the GP fit to the full data and to each of the approaches for combining 5 subsets. Visually each method does a reasonable job of capturing the full data fit, but the approximate methods are smoother and vary over a smaller range than if the full data had been used. The full data ranges between -1.9 and 3.1 whereas for the approximated fit using the divide and conquer approaches the range is -1.36 to 2.71 when combing with SwISS or consensus Monte Carlo algorithm. The barycentre approach varies over and shorter range of between

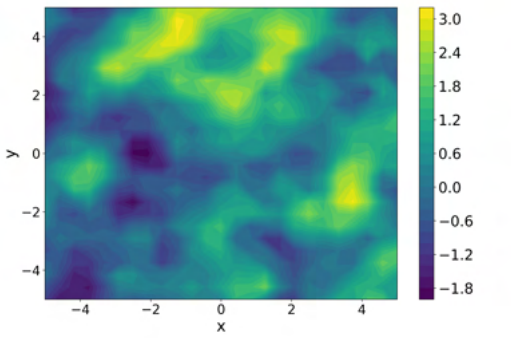


Figure 6.3.1: Simulated data over a grid.

-1.0 and 1.9. Therefore, the approximated methods are not capturing the higher and lower values as well as fitting the data to the full model. Figure 6.3.3 compares the standard deviations of the full data fit and the approximate methods, again for 5 subsets, compared to the full data fit the approximate methods over estimate the uncertainty. We consider the distribution plots of the values from Figures 6.3.2 and 6.3.3 to better compare how each method is performing. Figure 6.3.4 shows distribution plots of the expected value and standard deviation for each of the approximate methods and the full data fit. In both the 2 and 5 subset case the barycentre approach is failing to capture the tails. Whereas both the SwISS and consensus Monte Carlo algorithms only fail to capture the tails in the 5 subset case. This is not an issue in the 2 subset case shown in Figure 6.3.4a, implying the more we split the data the less accurate the approximate fit will be. We can also clearly see that SwISS is capturing the standard deviation better than the other approximate methods for both the 2 and 5 subset cases. Both consensus and the barycentre approach are resulting in much higher standard deviations. While SwISS is slightly higher it is capturing the full data fit standard deviation well. Comparing the Wasserstein distances in Table 6.3.1, as expected, SwISS is closest for the standard deviation in both cases at 0.01 and 0.03 for 2 and 5 subsets, respectively. The SwISS and consensus Monte Carlo algorithms both perform similarly for the expected value with a Wasserstein distance of 0.03 for 2 subsets, 0.09 for the Consensus Monte Carlo algorithm and 0.11 for SwISS for 5 subsets. The barycentre approach is notably

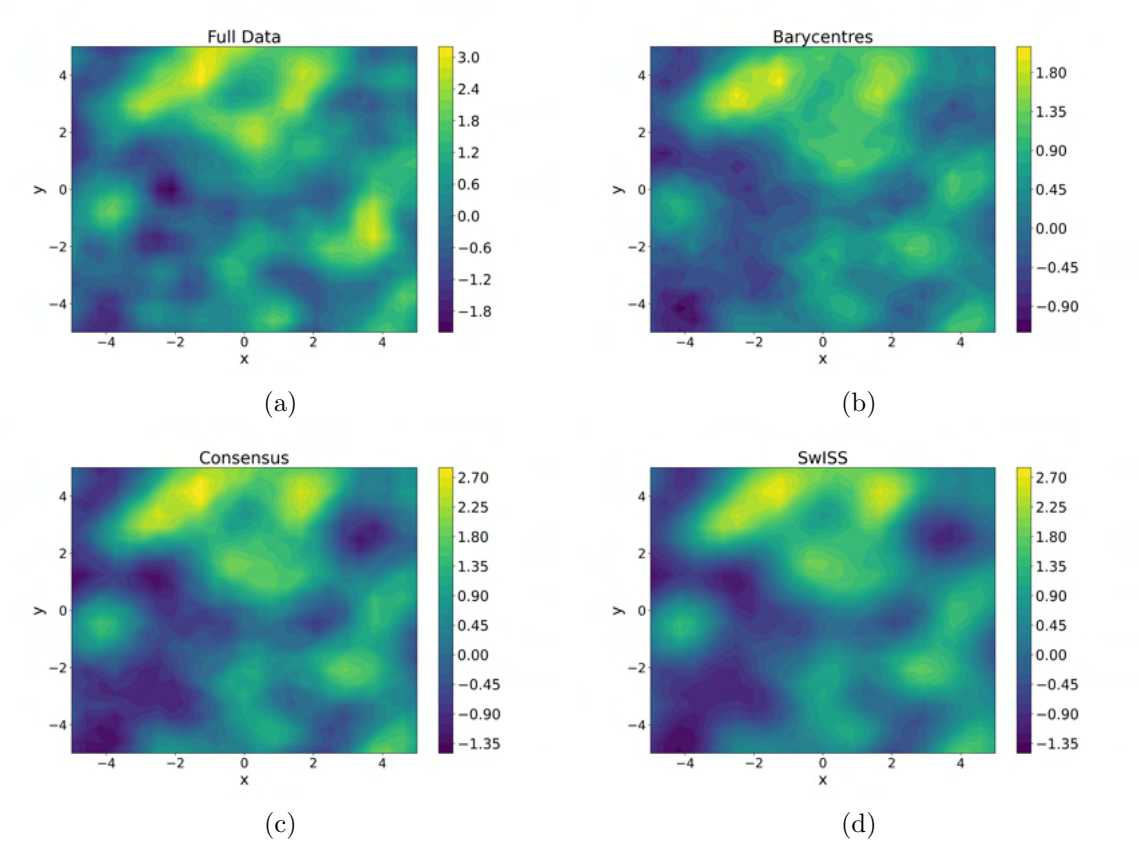
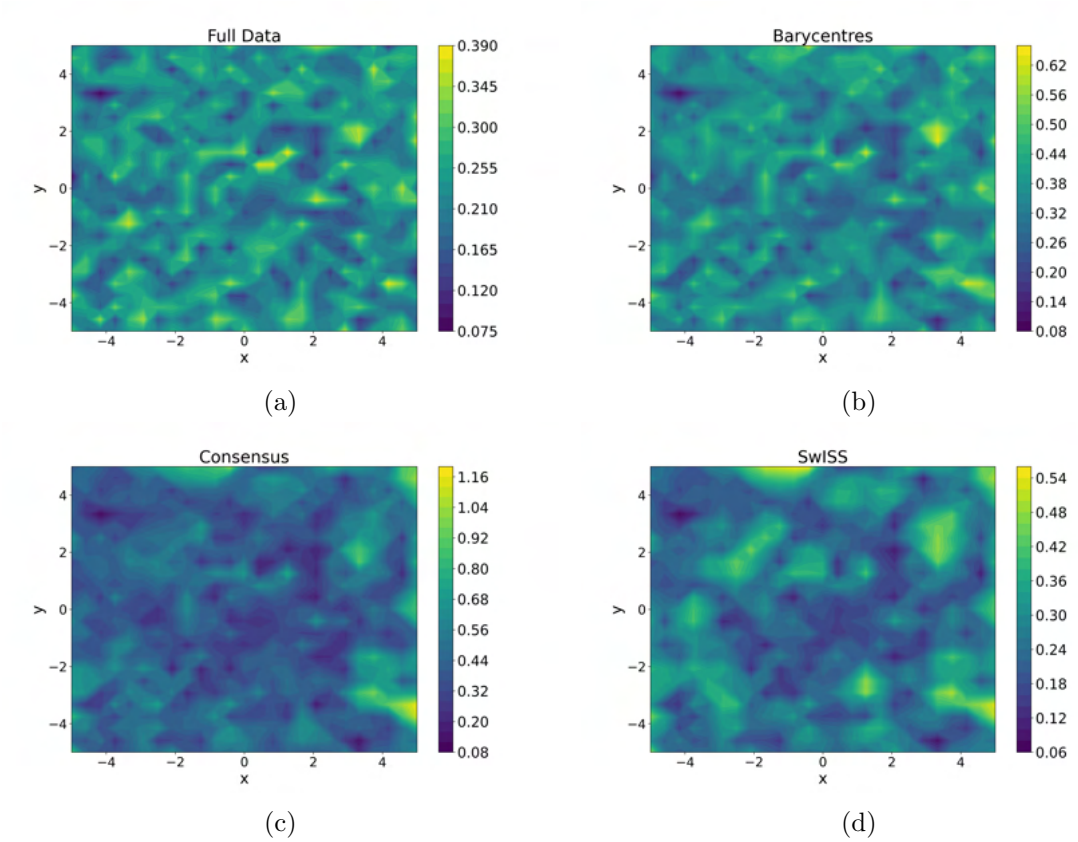


Figure 6.3.2: Expected values from the GP's predictive distribution for **(a)** the full data, and divide and conquer approaches using **(b)** Gaussian process barycentres, **(c)** consensus Monte Carlo and, **(d)** SwISS.

poorer with a Wasserstein distance of 0.19 and 0.34 for 2 and 5 subsets respectively.

When comparing our methods we consider 3 metrics, (i) the run time (ii) how well it captures the full data fit and (iii) the predictive error. We evaluate how well it captures the full data by considering the RMSE and R2 score between the expected values from the GP's predictive distribution for the full data approach and each of the approximate approaches. The predictive error is calculated by holding back 65 data points as a test portion of the data, the remaining 560 points are used to train the model. We compare the estimated values at the 65 unseen locations and compute the mean squared error. This is repeated 10 times and the average of these

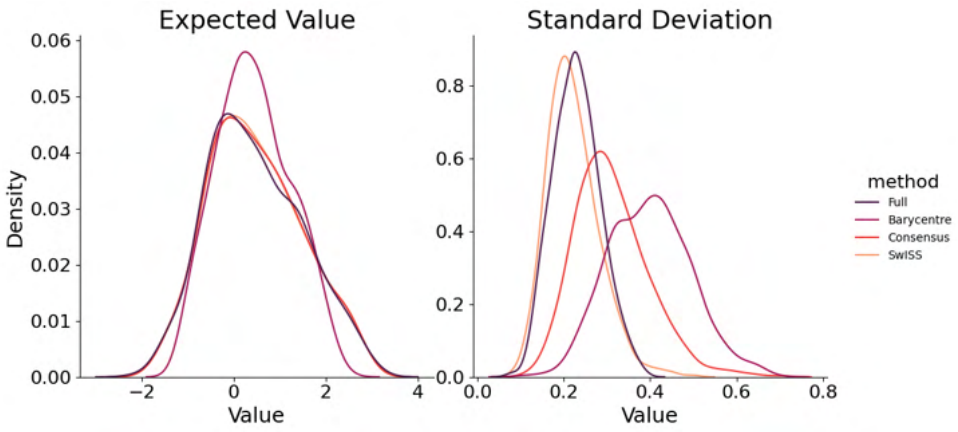


(a) (b) (c) (d)

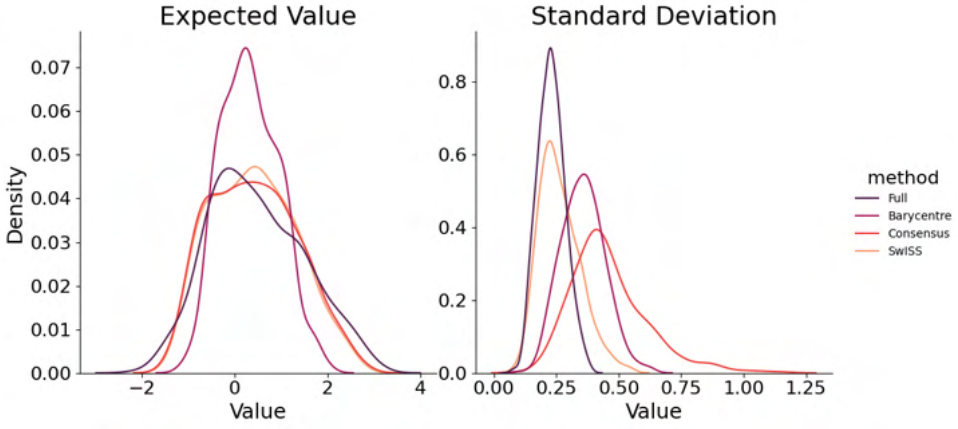
Figure 6.3.3: Standard deviation from the GP's predictive distribution for **(a)** the full data, and divide and conquer approaches using **(b)** Gaussian process barycentres, **(c)** consensus Monte Carlo and, **(d)** SwISS.

Table 6.3.1: Wasserstein distances for the expected values and standard deviations from the GP's predictive distribution of each of the approximate methods and the full data.

		Barycentre	**Consensus**	**SwISS**
2 Subsets	**Expected Value**	0.19	0.03	0.03
	Standard Deviation	0.17	0.08	0.01
5 Subsets	**Expected Value**	0.34	0.09	0.11
	Standard Deviation	0.13	0.23	0.03



(a)



(b)

Figure 6.3.4: Distribution plots of expected value and standard deviation for **(a)** 2 subsets and **(b)** 5 subsets of the data. The full data distribution plots are shown in both figures.

Table 6.3.2: Performance comparison of approximate methods compared to the full data for 2 and 5 subsets using simulated data.

	2 Subsets			**5 Subsets**			**Full Data**
	Barycentre	Consensus	SwISS	Barycentre	Consensus	SwISS	
MCMC Run Time	4.5 mins	4.6 mins	4.5 mins	2.6 mins	2.9 mins	2.6 mins	18.5 mins
Mean Subset Run Time	2.3 mins	2.3 mins	2.3 mins	0.52 mins	0.58 mins	0.52 mins	-
Combine Time	0.91 s	0.2 s	0.02 s	3.01 s	0.2 s	0.02 s	-
R2 Score	0.86	0.97	0.97	0.70	0.84	0.85	-
RMSE	0.37	0.17	0.17	0.57	0.41	0.40	-
Prediction Error	0.3	0.16	0.15	0.55	0.33	0.32	0.09

10 runs is the predictive error. Table 6.3.2 summarises the comparison between the methods for 2 and 5 subsets. Fitting to the full data takes 18.5 minutes, the SwISS algorithm is slightly faster than the other methods at 4.5 minutes and 2.6 minutes for 2 and 5 subsets respectively. Consensus Monte Carlo is similar at 4.6 minutes and 2.9 minutes for 2 and 5 subsets respectively. The Barycentre approach has a longer combining time compared to SwISS and consensus Monte Carlo, and this combining time increases with the number of subsets, 0.91 seconds for 2 subsets and 3.1 seconds for 5 subsets. Therefore, even though the barycentre approach fits in the same time as the SwISS algorithm it does take slightly longer due to the combining time. When compared to the fit of the full data for the 2 subset case, both the SwISS and consensus Monte Carlo algorithms achieve a similar R2 score of 0.97 and RMSE of 0.17. The barycentre approach performs worse at 0.86 and 0.37 for the R2 score and RMSE respectively.The predictive error when fitting to the full data is 0.09, the barycentre approach has the highest prediction error at 0.30 compared to 0.16 and 0.15 for consensus Monte Carlo and SwISS, respectively. For the 5 subset case, when compared to the fit of the full data, the Barycentre approach performs the poorest with an R2 score of 0.70 and RMSE 0.57. Both the Consensus Monte Carlo and SwISS algorithms performed similarly with R2 scores of 0.84 and 0.85, respectively, and RMSE of 0.41 and 0.40, respectively. The Barycentre approach again performs worse with a prediction error of 0.55 and Consensus Monte Carlo and SwISS algorithms performed similarly with a prediction error of 0.33 and 0.32. Overall, SwISS preformed marginally better in terms of computational speed and accuracy compared to the consensus Monte Carlo algorithm and thus could be seen as the preferred choice. Also, as previously discussed, SwISS captures the standard deviation better than the other methods. However, compared to the SwISS and consensus Monte Carlo algorithms the barycentre approach fails to capture the full data fit as well as the other methods.

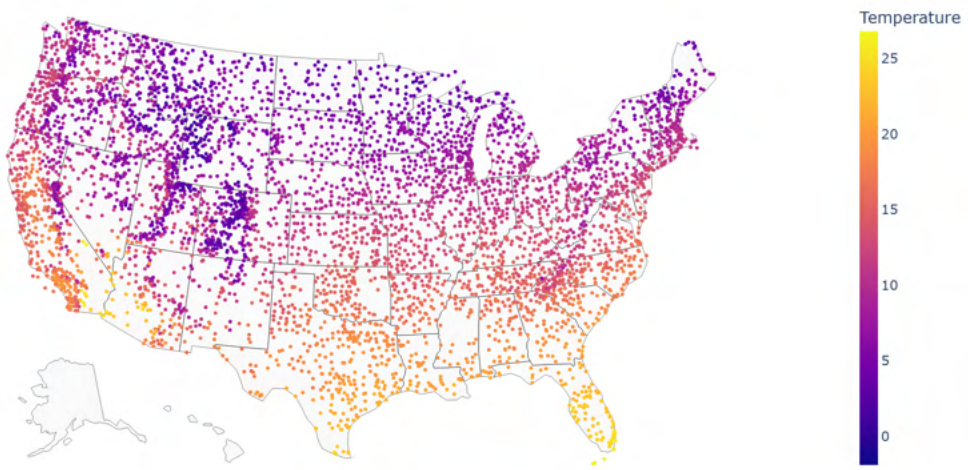


Figure 6.3.5: Map of 2018 annual average temperature in USA.

6.3.2 Real data study: USA Temperature Data

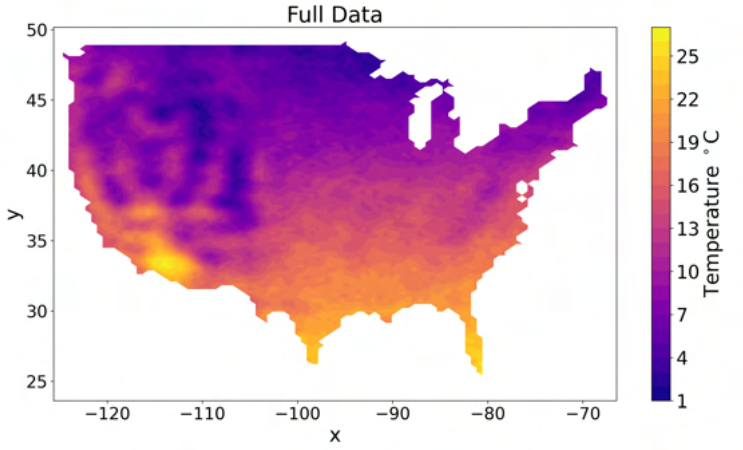
To demonstrate the accuracy of the combination methods using real data, we consider temperature data at measurement sites in the USA. Data is taken from the National Centers for Environmental Information (https://www.ncei.noaa.gov/access). Due to our focus on spatial modelling, we can ignore the temporal component and we will consider annual average temperature for 2018. After removing any sites with missing data we are left with 5660 measurement sites. While fitting a GP to over 5000 data points is feasible it will still have a considerable run time, thus this is an appropriate size of dataset to implement these methods on. Figure 6.3.5 shows the locations of the monitoring stations, the data is fairly smooth with warmer temperatures in the south and lower temperatures in the north. There are some lower temperatures to the west of the map, which can be associated with a higher elevation at these locations. However, in the interest of simplicity in the model, and retaining the focus of the combining aspect of this work, we will not be including elevation data in the model.

As before, we will compare the full data fit to the approximate methods using

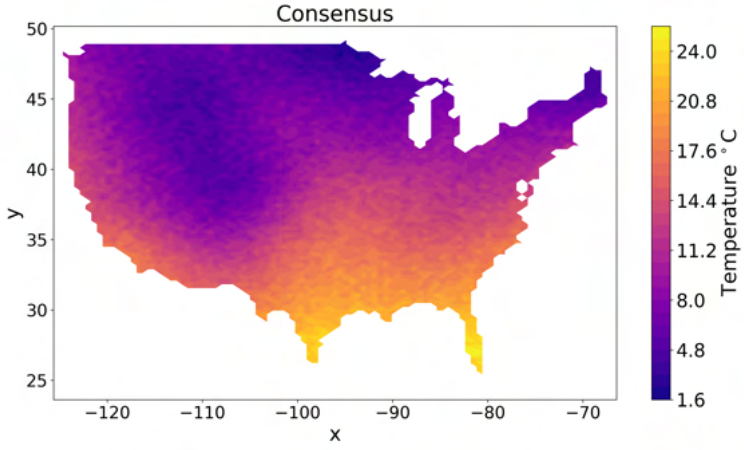
the same criteria as before. Due to the larger, dataset we will split the data into 5 and 10 equal subsets, again randomly sampled and non-overlapping. We standardise the data by subtracting the mean and dividing by the standard deviation to avoid numerical instabilities when fitting the GP. Following on from the results of the previous section we will focus on the consensus Monte Carlo and SwISS algorithms. Combining the posteriors using barycentres is less accurate than the chosen methods.

Figure 6.3.6 shows the expected value for the full data fit and the approximate methods using the consensus Monte Carlo and SwISS algorithms for 10 subsets. Visually we can see that the approximate methods are much smoother that the full data fit. We can also note that the range of values is shorter for the approximate methods, between 1.6 °C and 25 °C for consensus Monte Carlo and 4 °C and 23.5 °C for SwISS compared to between 1 °C and 26 °C for the full data. This implies that consensus Monte Carlo is performing better at capturing the range of values. We can further see this in Figure 6.3.10b when comparing the distribution plots of the expected values at each location; the tails for SwISS do not capture the tails of the full data distribution. Now considering the 5 subset case in Figure 6.3.7, consensus Monte Carlo is less smooth and able to capture more of the features than the full data fit has and both consensus Monte Carlo and SwISS show a larger range of values with SwISS overestimating the temperature in places. This matches with Figure 6.3.10a, where the tails of each of the methods are much closer together.

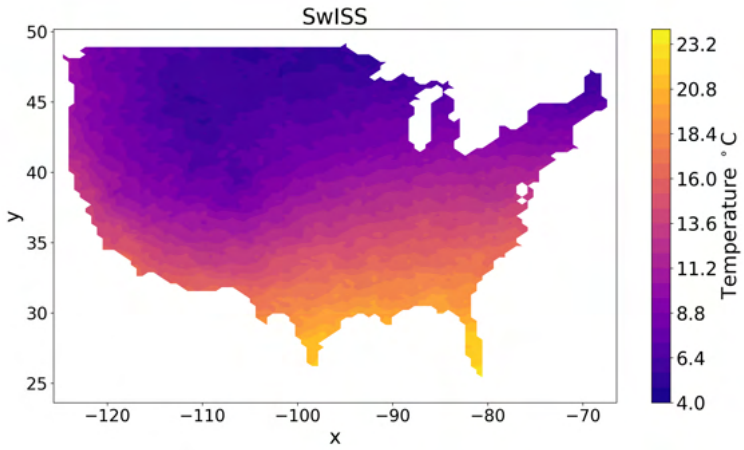
Figure 6.3.8 shows the standard deviation for the full data fit and the approximate methods using consensus Monte Carlo and SwISS for 10 subsets. We can also see that the range of values is larger for the approximate methods, between 0.45 °C and 4.05 °C for consensus Monte Carlo and 0.4 °C and 5.3 °C for SwISS compared to between 0.3 °C and 3.0 °C for the full data. This implies both methods are overestimating the standard deviation. We can further see this in Figure 6.3.10



(a)

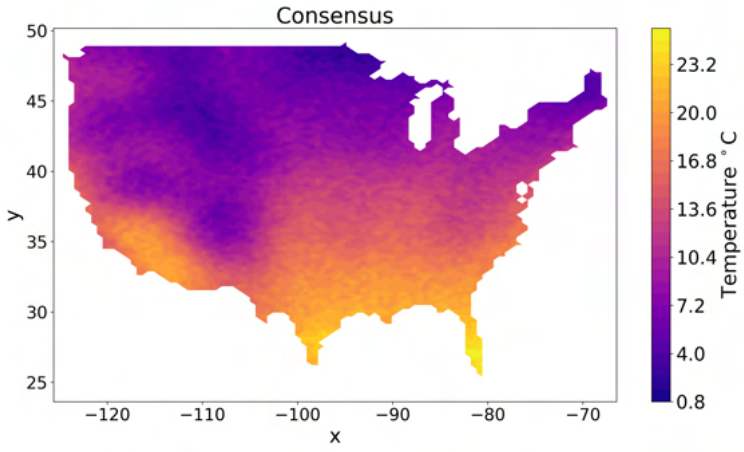


(b)

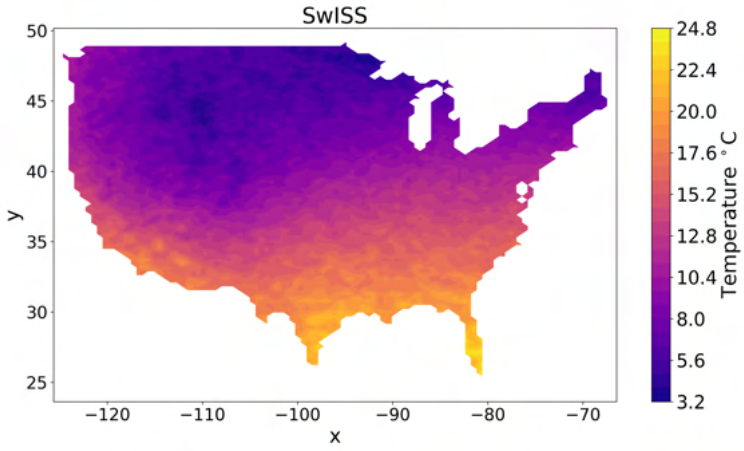


(c)

Figure 6.3.6: Expected values from the GP's predictive distribution for **(a)** the full data, and divide and conquer approaches using **(b)** consensus Monte Carlo and, **(c)** SwISS for 10 subsets.

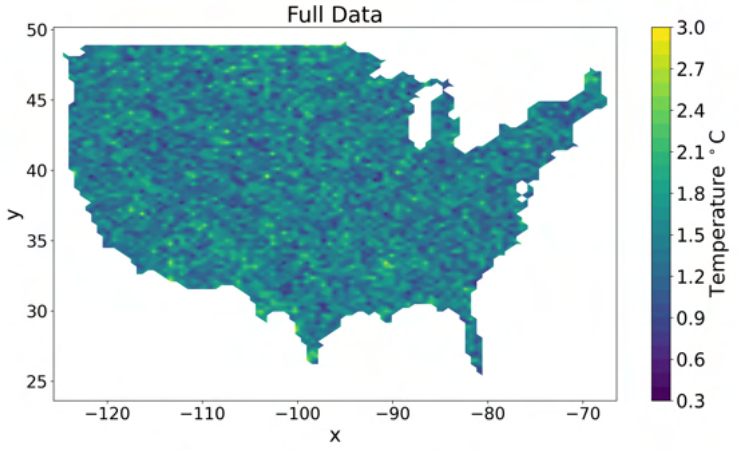


(a)

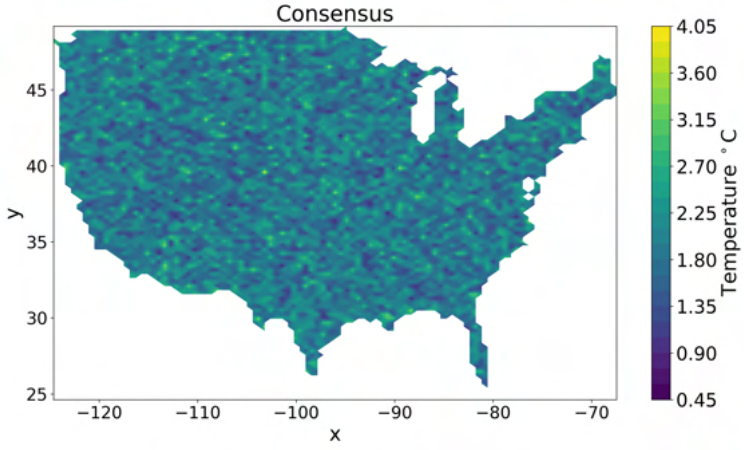


(b)

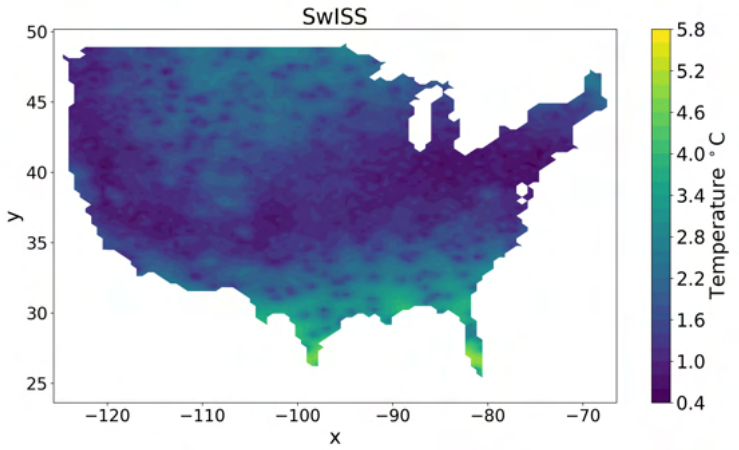
Figure 6.3.7: Expected values from the GP's predictive distribution for divide and conquer approaches using **(a)** consensus Monte Carlo and, **(b)** SwISS for 5 subsets.



(a)

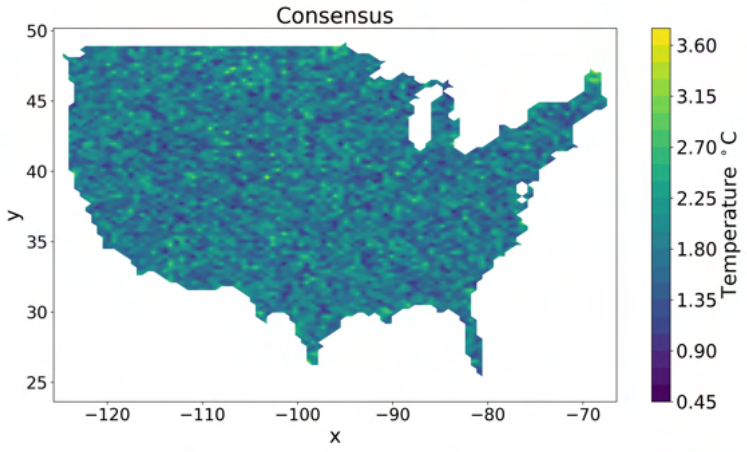


(b)

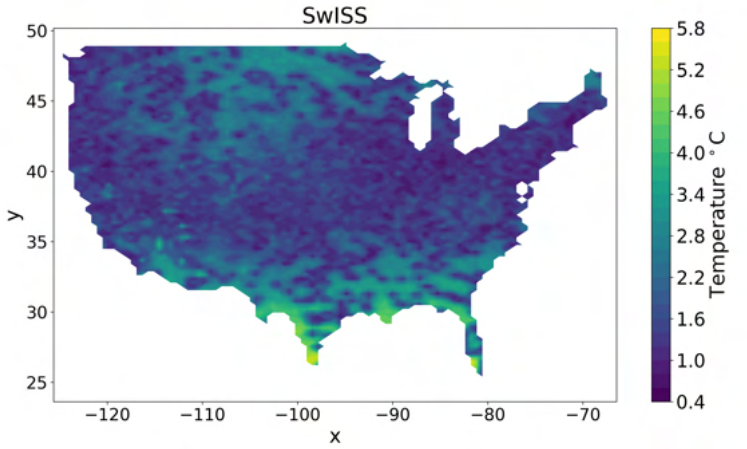


(c)

Figure 6.3.8: Standard deviations from the GP's predictive distribution for **(a)** the full data, and divide and conquer approaches using **(b)** consensus Monte Carlo and, **(c)** SwISS.

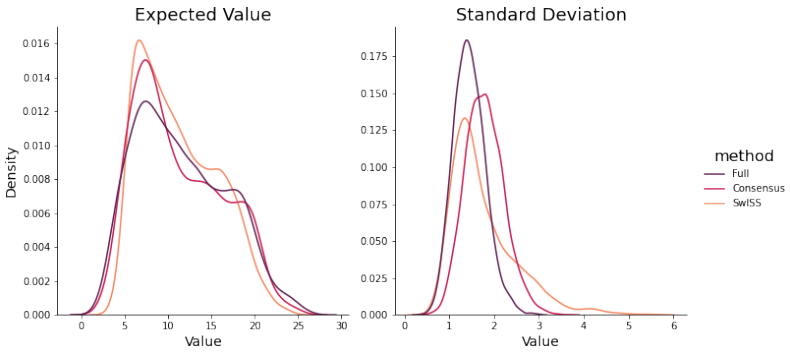


(a)

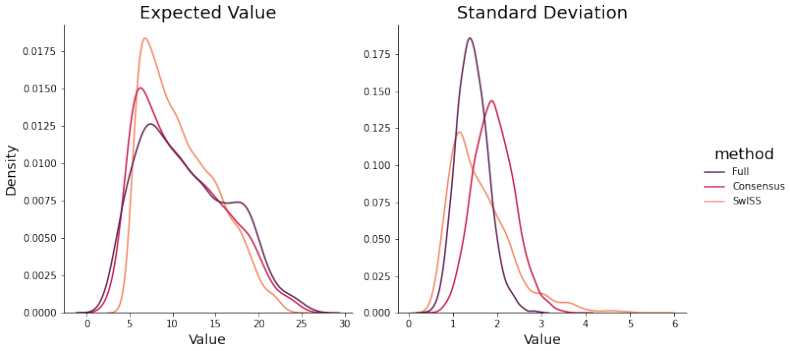


(b)

Figure 6.3.9: Standard deviations from the GP's predictive distribution for divide and conquer approaches using **(a)** consensus Monte Carlo and, **(c)** SwISS.



(a)



(b)

Figure 6.3.10: Distribution plots of expected value and standard deviation for **(a)** 5 subsets and **(b)** 10 subsets of the data. The full data distribution plots are shown in both figures.

when comparing the distribution plots of the standard deviation at each location SwISS does a better job than consensus Monte Carlo for capturing the lower values of standard deviation. However, SwISS has a much longer tail to the right leading to higher overestimates compared to SwISS. Looking at the 5 subset case in Figure 6.3.9, we have a smaller ranger over which the uncertainties are varying. We can see in Figure 6.3.10a the approximate methods are still overestimating the uncertainty and not capturing the lower values.

Table 6.3.3 shows the Wasserstein distances for the distribution plots in Figure 6.3.10. When comparing the expected values to the full data fit we have a Wasserstein distance of 0.63 and 1.07 for consensus Monte Carlo and SwISS, respectively, for 10 subsets. Thus, we can conclude consensus Monte Carlo better captures the expected value compared to the full data fit. Conversely, SwISS better

Table 6.3.3: Wasserstein distances for the expected value and standard deviation of each of the approximate methods and the full data for USA temperature data for 5 and 10 subsets.

	5 subsets		**10 subsets**	
	Consensus	**SwISS**	**Consensus**	**SwISS**
Expected Value	0.32	0.69	0.63	1.07
Standard Deviation	1.66	1.63	1.80	1.50

Table 6.3.4: Performance comparison of approximate methods compared to the full data for 5 and 10 subsets using USA temperature data.

	5 subsets		**10 subsets**		
	Consensus	**SwISS**	**Consensus**	**SwISS**	**Full Data**
MCMC Run Time	6.6 hrs	6.4 hrs	3.8 hrs	3.7 hrs	2 days
Subset Run Time	79 mins	77 mins	23 mins	22 mins	-
R2 Score	0.94	0.90	0.88	0.86	-
RMSE	1.32 °C	1.62 °C	1.82 °C	1.96 °C	-

captures the standard deviation with a Wasserstein distance of 1.50 compared to 1.80 for consensus Monte Carlo for 10 subsets. The ability of SwISS to capture the standard deviation better aligns with our previous results in the simulated data study. The results are similar for 5 subsets with consensus having a lower Wasserstein distance for the expected value, 0.32 compared to 0.69 for SwISS. SwISS had a slightly lower Wasserstein distance of 1.63 compared to 1.66 for consensus Monte Carlo.

As before, we evaluate our methods by considering the runtime and how well they capture the full data. Table 6.3.4 summaries our results, the full data takes around 2 days to fit, both SwISS and consensus Monte Carlo fit in a similar amount of time with SwISS being slightly faster in 3.7 hours compared to 3.8 hours for consensus Monte Carlo. Both approximate methods have a similar R2 score 0.86 and 0.88 for consensus Monte Carlo and SwISS, respectively. Again, we compare how well it captures the full data by considering the RMSE and R2 score between the expected values from the GP's predictive distribution for the full data approach and the approximate methods. Consensus Monte Carlo has the lower RMSE at 1.82 °C compared to 1.96 °C for SwISS for the 10 subset case. Again, for the 5 subset case they fit in a similar time with SwISS being slightly faster at 6.4 hours, consensus Monte Carlo has a higher R2 score of 0.94 compared to SwISS at 0.90 and consensus Monte Carlo has a lower RMSE at 1.32 °C compared to SwISS at 1.62 °C.

6.4 Discussion and Conclusion

Dividing the data into more subsets is more computationally efficient, but the approximate methods perform worse as the number of subsets increases. Sub-setting the data also risks losing information when recombining, this was noticeable in the real data case where the approximate methods were noticeably smoother. This could be addressed through additional covariates such as elevation which we did not

consider here.

The barycentre method performed the poorest compared to the other methods. SwISS performed best at capturing the standard deviation in both the simulated and real data. For the simulated data, SwISS and Consensus performed similarly at capturing the expected value. Consensus Monte Carlo performed slightly better for the real data with 5 and 10 subsets. However, SwISS was still better at capturing the standard deviation.

Further work could look at extending these methods to produce an exact result, as well as considering methods of best practice when sub-setting the data. Here we randomly sampled from the data, but depending on the application it may make more sense to subset by region, or another criterion related to the application.

Chapter 7

Conclusions

The aim of this thesis was to contribute to methods for efficient handling of environmental data, addressing the common issue of incomplete data, with applications looking at missing time-series data and a large data spatial problem with unseen locations. This thesis is concluded by summarising the main results of this work and discussing the limitations of this work and possible ways to address them. Finally, this thesis presents potential future extensions to this work.

7.1 Summary of Main Results

This thesis proposed an extension to the traditional Kalman filter that provides a computationally efficient method for bias-correcting data by conceptualising the bias as a skew between the biased data set and a second dataset that is assumed to be unbiased. Further, this thesis has shown there are identification issues with the unified skew-normal distribution and computational issues using this distribution for the skew Kalman filter. Additionally, this thesis demonstrated the feasibility of reducing the skew Kalman filter to a multivariate skew normal and implemented an efficient method to accurately estimate the parameters. In Chapter 4 we derived a new simplified approach to the skew Kalman filter currently proposed in the literature which allows us to retain the simplicity and computational efficiency that

the traditional Kalman filter is known for. This was achieved by deriving the skew Kalman filter using a multivariate skew-normal distribution instead of the unified skew-normal distribution. To address the difficulty of estimating the parameters we proposed a two-step approach using the unbiased dataset to estimate the skew Kalman filter parameters that insured the computational efficiency of the method was retained.

Extending this work, this thesis demonstrated how to implement the skew Kalman filter as a tool for bias correction and missing data. Chapter 5 implemented the skew Kalman filter approach to use the bias corrected data to infill missing data in the unbiased dataset. This method allowed for a relatively simple approach to infilling missing data that had a short computational time and was effective at infilling the missing data. The method was demonstrated using surface level ozone data and shown to perform well for data which had a combination of randomly and consecutively missing data in the real world scenario presented. As this method is not modelling the ozone data it would be straight forward to implement in a range of different applications provided a suitable secondary bias dataset was available. Using a real world scenario of missing data, the skew Kalman approach out performed the compared methods for estimating the missing data.

Further, this thesis presents the necessary foundations for implementing divide-and-conqueror approaches for spatial data by evaluating existing methods from the independent setting in Chapter 6. While divide-and-conquer methods are popular in the independent setting, this thesis explored their effectiveness in the spatial setting where assumptions of independence no longer hold. Three methods are compared and evaluated based on computationally efficiency and how well they capture the full data fit. It has been shown the both the consensus Monte Carlo algorithm and SwISS can be used in the spatial setting to approximate the full data fit for large datasets, significantly reducing the computational time compared to using the full

data.

7.2 Future Work

The work presented in this thesis can be further extended in ways that are both interesting from a methods point of view and an application focus. Despite the difficulties for parameter estimation in the skew Kalman filter, Kalman filtering for non-Gaussian data is still an interesting field of research. In Chapter 4, we reduced the skew Kalman filter to follow a multivariate skew normal distribution and this thesis proposed a 2-step approach for estimating the parameters when the underlying signal was known. This decision was made to preserve the simplicity and computational efficiency of the Kalman filter and while there are still situations where this would be useful it is naturally limiting for this to be the case. One of the main reasons for reducing the filter to the multivariate skew normal was the computational challenges in estimating parameters using the unified skew-normal distribution. However, this results in the filtering equations that do not have a closed form and as such one of the filtering equations is incalculable. While this was not a problem for the implementation used in this thesis as the incalculable equation did not feature in the likelihood, a closed-form version would be more robust.

Traditionally, extensions of the multivariate skew-normal distribution such as the closed skew normal or unified skew normal are used in the skew Kalman filter, as these results in filtering equations with a closed form. However, working with the unified skew normal or closed skew normal in the skew Kalman filter is computationally challenging as there are multivariate normal cdf terms in the pdf which increase in dimension with each time step and evaluating high-dimensional multivariate normal cdfs is very expensive computationally. One possible option would be to address this growing dimension problem, recent work by (Guljanov et al. 2022) proposes a pruning algorithm to the updating step of the skew Kalman

filter to overcome the increasing dimension issue.

The work in Chapter 4 also highlighted issues in parameter identification using the unified skew normal, and Guljanov et al. (2022) refer to identification issues with the closed skew normal also. Addressing these identification issues in these distributions would be a sensible next step for the skew Kalman filter, recent work by Wang et al. (2023) explored some of the issues of the non-identifiability of the unified skew-normal distribution. This work discussed some identifiable sub-models of the unified skew normal and it may be possible to derive a closed-form version of the skew Kalman filter using these that would not suffer from the identification issues found in the unified skew normal. Another proposed solution is fixing some of the terms of the unified skew normal, in our approach in Chapter 4 we fixed some of the parameters of the unified skew normal such that it reduced to the multivariate skew normal, alternatively, we could fix certain parameters through the filtering steps but retain the unified skew normal distribution instead of reducing it to the multivariate skew-normal.

Next, looking at the skew Kalman filter approach for missing data, aside from implementing the skew Kalman filter with the methods discussed above, possible next steps for this model would be improving how the skewness between the datasets is estimated in the presence of missing data. In Chapter 5 a seasonal approach is used to estimate the skewness, taking the average skewness from other years. While this is suitable in the air quality setting, as many air pollutants follow an annual cycle, the generalisability of the model could be improved by further refining this. If the skew is relatively constant a mean skewness could be used, or another option would be to also model the skewness between the datasets. The model could be further refined by incorporating additional information from the dataset that has the missing dataset to further improve the estimate for the missing data. This could be achieved by combing the skew Kalman filter with another missing data method,

such as one of the methods discussed in Chapter 2, and using a weighting between them.

An immediate extension to the divide-and-conquer work in Chapter 6 would be to achieve an exact posterior for the combined subsets compared to the approximate posterior currently presented. This could be achieved using an importance sampling technique (Geweke 1989) where the results from the combining strategy could be used as the proposal and the weighting between this and the full data posterior used to achieve an exact result.

While this work primarily focussed on how to combine the subsetted data, another aspect of this approach is how the data is subsetted to begin with. Alternatives could include splitting the data regionally as this could be most practical depending on the application. One possible method to achieve this would be to use Gaussian Markov random fields (Rue & Held 2005), which is an undirected graphical model where each edge represents a dependency. Representing the data with this structure could allow for more natural subsetting of the data depending on the application domain. Another consideration when subsetting the data is that if there are features present in only a small number of locations, ensuring that information is suitably preserved so that it is not lost in the combining stage.

Appendix A

Appendix

A.1 Properties of the Gaussian distribution

Definition A.1.1. *(Gaussian distribution) A random variable $x \in \mathbb{R}^n$ has a Gaussian distribution with mean $m \in \mathbb{R}^n$ and covariance $P \in \mathbb{R}^{n \times n}$ if its probability density function has the form Särkkä (2013)*

$$N(x|m,P) = \frac{1}{(2\pi)^{n/2}\det(P)^{1/2}} \exp -\frac{1}{2}(x-m)^T P^{-1}(x-m). \tag{A.1.1}$$

Lemma A.1.1. *(Joint distribution of Gaussian variables) If random variable $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ have the Gaussian probability distributions (Särkkä 2013)*

$$\begin{aligned} x &\sim N(m,P) \\ y|x &\sim N(Hx+u,R), \end{aligned} \tag{A.1.2}$$

then the joint distribution of x, y is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} m \\ Hm+u \end{pmatrix}, \begin{pmatrix} P & PH^T \\ HP & HPH^T+R \end{pmatrix} \right), \tag{A.1.3}$$

and the marginal distribution of y is

$$y \sim N(Hm + u, HPH^T + R). \tag{A.1.4}$$

Lemma A.1.2. *(Conditional distribution of Gaussian variable) If the random variables x and y have the joint Gaussian probability distribution*

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \right), \tag{A.1.5}$$

where as before $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$ and the dimensions of the mean vectors and covariance matrix sub-blocks are chosen to match x and y. Then the marginal and conditional distributions of x and y are (Särkkä 2013)

$$\begin{aligned} x &\sim N(\mu_x, \sigma_{xx}) \\ y &\sim N(\mu_y, \sigma_{yy}) \\ x|y &\sim N(\mu_x + \sigma_{xy}\sigma_{yy}^{-1}(y - \mu_y), \sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx}) \\ y|x &\sim N(\mu_y + \sigma_{yx}\sigma_{xx}^{-1}(x - \mu_x), \sigma_{yy} - \sigma_{yx}\sigma_{xx}^{-1}\sigma_{xy}) \end{aligned} \tag{A.1.6}$$

Proof. (Conditional of a joint Gaussian is Gaussian)
We want to show conditional densities given by

$$\begin{aligned} p(x|y) &= \frac{p(x, y; \mu, \sigma)}{\int_{x \in \mathbb{R}^n} p(x, y; \mu, \sigma) dx} \\ p(y|x) &= \frac{p(x, y; \mu, \sigma)}{\int_{y \in \mathbb{R}^m} p(x, y; \mu, \sigma) dy} \end{aligned} \tag{A.1.7}$$

are also Gaussian such that

$$\begin{aligned} x|y &\sim N(\mu_x + \sigma_{xy}\sigma_{yy}^{-1}(y - \mu_y), \sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx}) \\ y|x &\sim N(\mu_y + \sigma_{yx}\sigma_{xx}^{-1}(x - \mu_x), \sigma_{yy} - \sigma_{yx}\sigma_{xx}^{-1}\sigma_{xy}). \end{aligned} \tag{A.1.8}$$

First we write the conditional density explicitly

$$\begin{aligned} p(y|x) &= \frac{p(x,y;\mu,\sigma)}{\int_{y\in\mathbb{R}^m} p(x,y;\mu,\sigma)dy} \\ &= \frac{1}{Z}\exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ \\ y-\mu_y \end{bmatrix}^T \begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix}^{-1} \begin{bmatrix} x-\mu_x \\ \\ y-\mu_y \end{bmatrix}\right) \end{aligned} \tag{A.1.9}$$

where Z is a normalisation constant containing any terms that do not depend on y. Next we rewrite our inverse covariance matrix σ^{-1} such that

$$\sigma^{-1} = V = \begin{bmatrix} V_{xx} & V_{xy} \\ \\ V_{yx} & V_{yy} \end{bmatrix} \tag{A.1.10}$$

then,

$$\begin{aligned} p(y|x) &= \frac{1}{Z}\exp\left(-\frac{1}{2}\begin{bmatrix} x-\mu_x \\ \\ y-\mu_y \end{bmatrix}^T \begin{bmatrix} V_{xx} & V_{xy} \\ \\ V_{yx} & V_{yy} \end{bmatrix} \begin{bmatrix} x-\mu_x \\ \\ y-\mu_y \end{bmatrix}\right) \\ p(y|x) &= \frac{1}{Z}\exp\Bigg(-\Bigg[\frac{1}{2}(x-\mu_x)^T V_{xx}(x-\mu_x) \\ &\qquad + \frac{1}{2}(x-\mu_x)^T V_{xy}(y-\mu_y) \\ &\qquad + \frac{1}{2}(y-\mu_y)^T V_{yx}(x-\mu_x) \\ &\qquad + \frac{1}{2}(y-\mu_y)^T V_{yy}(y-\mu_y)\Bigg]\Bigg) \end{aligned} \tag{A.1.11}$$

Next we look to use the completion of squares method. Consider the quadratic function

$$z^T A z + b^T z + c \tag{A.1.12}$$

where A is a symmetric and non singular matrix. Then,

$$z^T Az + b^T z + c = \frac{1}{2}(z + A^{-1}b)^T A(z + A^{-1}b) + c - \frac{1}{2}b^T A^{-1}b. \tag{A.1.13}$$

To apply completion of squares we let

$$\begin{aligned} z &= y - \mu_y \\ A &= V_y y \\ b &= V_{yx}(x - \mu_x) \\ c &= \frac{1}{2}(x - \mu_x)^T V_{xx}(x - \mu_x) \end{aligned} \tag{A.1.14}$$

and we can now rewrite Equation (A.1.11) as follows

$$\begin{aligned} p(y|x) = \frac{1}{Z}\exp\Bigg(- \Bigg[&\frac{1}{2}(y - \mu_y + V_{yy}^{-1}V_{yx}(x - \mu_x))^T V_{yy}(y - \mu_y + V_{yx}^{-1}(x - \mu_x)) \\ &+ \frac{1}{2}(x - \mu_x)^T V_{xx}(x - \mu_x) - \frac{1}{2}(x - \mu_x)^T V_{xy}V_{yy}^{-1}V_{yx}(x - \mu_x) \Bigg] \Bigg) \end{aligned} \tag{A.1.15}$$

pulling all the terms that do not depend on y into a new normalisation constant gives.

$$p(y|x) = \frac{1}{Z'}\exp\Bigg(- \Bigg[\frac{1}{2}(y - \mu_y + V_{yy}^{-1}V_{yx}(x - \mu_x))^T V_{yy}(y - \mu_y + V_{yx}^{-1}(x - \mu_x)) \Bigg] \Bigg) \tag{A.1.16}$$

Equation (A.1.16) has the form of a Gaussian density with mean $\mu_y - V_{yy}^{-}1V_{yx}(x - \mu_x)$ and covariance V_{yy}^{-1}. To show this is the same as Equation (A.1.8) we have

$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} (V_{xx} - V_{xy}V_{yy}^{-1}V_{yx})^{-1} & -(V_{xx} - V_{xy}V_{yy}^{-1}V_{yx})^{-1}V_{xy}V_{yy}^{-1} \\ -V_{yy}^{-1}V_{yx}(V_{xx} - V_{xy}V_{yy}^{-1}V_{yx})^{-1} & (V_{yy} - V_{yx}V_{xx}^{-1}V_{xy})^{-1} \end{bmatrix} \tag{A.1.17}$$

which gives

$$\mu_{y|x} = \mu_y - V_{yy}^{-1} V_{yx}(x - \mu_x) = \mu_y - \sigma_{yx}\sigma_{xx}^{-1}(x - \mu_x). \tag{A.1.18}$$

Conversely, we have

$$\begin{bmatrix} V_{xx} & V_{xy} \\ V_{yx} & V_{yy} \end{bmatrix} = \begin{bmatrix} (\sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx})^{-1} & -(\sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx})^{-1}\sigma_{xy}\sigma_{yy}^{-1} \\ -\sigma_{yy}^{-1}\sigma_{yx}(\sigma_{xx} - \sigma_{xy}\sigma_{yy}^{-1}\sigma_{yx})^{-1} & (\sigma_{yy} - \sigma_{yx}\sigma_{xx}^{-1}\sigma_{xy})^{-1} \end{bmatrix} \tag{A.1.19}$$

which gives

$$\sigma_{y|x} = V_{yy}^{-1} = \sigma_{yy} - \sigma_{yx}\sigma_{xx}^{-1}\sigma_{xy}. \tag{A.1.20}$$

□

Bibliography

Agudelo, O. M., Barrero, O., Péter, V. & Moor, B. D. (2011), 'Assimilation of ozone measurements in the air quality model aurora by using the ensemble kalman filter', *Proceedings of the IEEE Conference on Decision and Control* pp. 4430–4435.

Allison, P. D. (2001), *Missing Data*, SAGE Publications.

Arellano-Valle, R. B. & Azzalini, A. (2006), 'On the unification of families of skew-normal distributions', *Scandinavian Journal of Statistics* **33**, 561–574.

Arellano-Valle, R. B. & Azzalini, A. (2020), 'Some properties of the unified skew-normal distribution'.

Arellano-Valle, R. B., Contreras-Reyes, J. E., Quintero, F. O. & Valdebenito, A. (2019), 'A skew-normal dynamic linear model and bayesian forecasting', *Computational Statistics* **34**, 1055–1085.

Arellano-Valle, R. B. & Genton, M. G. (2005), 'On fundamental skew distributions', *Journal of Multivariate Analysis* **96**, 93–116.
URL: *www.elsevier.com/locate/jmva*

Baraldi, A. N. & Enders, C. K. (2010), 'An introduction to modern missing data analyses', *Journal of School Psychology* **48**, 5–37.

Baruah, B., Dutta, M. P. & Bhattacharyya, D. K. (2023), 'An effective ensemble method for missing data imputation', *International Journal of Information and*

Computer Security **20**, 295.
URL: *http://www.inderscience.com/link.php?id=128846*

Braak, C. J. F. T., Strien, A. J. V., Meder, R. & Verstrael, T. J. (1994), 'Analysis of monitoring data with many missing valves: Whlch method?', *Statistics Netherlands* pp. 663–673.

Buhrmester, V., Münch, D. & Arens, M. (2021), 'Analysis of explainers of black box deep neural networks for computer vision: A survey', *Machine Learning and Knowledge Extraction* **3**, 966–989.

Carvalho, C. G. N. D., Gomes, D. G., Souza, J. N. D. & Agoulmine, N. (2011), 'Multiple linear regression to improve prediction accuracy in wsn data reduction'.

Casciaro, G., Cavaiola, M. & Mazzino, A. (2022), 'Highlights calibrating the cams european multi-model air quality forecasts for regional air pollution monitoring calibrating the cams european multi-model air quality forecasts for regional air pollution monitoring'.

Choudhury, S. J. & Pal, N. R. (2019), 'Imputation of missing data with neural networks for classification', *Knowledge-Based Systems* **182**, 104838.
URL: *https://linkinghub.elsevier.com/retrieve/pii/S0950705119303132*

Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L., Pope, C. A., Shin, H., Straif, K., Shaddick, G., Thomas, M., van Dingenen, R., van Donkelaar, A., Vos, T., Murray, C. J. & Forouzanfar, M. H. (2017), 'Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015', *The Lancet* **389**, 1907–1918.
URL: *http://dx.doi.org/10.1016/S0140-6736(17)30505-6*

Coulibaly, P. & Evora, N. D. (2007), 'Comparison of neural network methods for infilling missing daily weather records', *Journal of Hydrology* **341**, 27–41.

Cover, T. & Hart, P. (1967), 'Nearest neighbor pattern classification', *IEEE transactions on information theory* **13**, 21–27.

Cremers, J. (2020), 'waspr package - rdocumentation'.
URL: *https://www.rdocumentation.org/packages/waspr/versions/1.0.0*

Dee, D. P., Balmaseda, M., Balsamo, G., Engelen, R., Simmons, A. J. & Thépaut, J.-N. (2014), 'Toward a consistent reanalysis of the climate system', *Bulletin of the American Meteorological Society* **95**, 1235–1248.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the em algorithm', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**, 1–22.

Djalalova, I., Monache, L. D. & Wilczak, J. (2015), 'Corrigendum to pm2.5 analog forecast and kalman filter post-processing for the community multiscale air quality (cmaq) model', *Atmospheric Environment* **119**, 431–442.
URL: *http://dx.doi.org/10.1016/j.atmosenv.2015.05.057*

Donders, A. R. T., van der Heijden, G. J., Stijnen, T. & Moons, K. G. (2006), 'Review: A gentle introduction to imputation of missing values', *Journal of Clinical Epidemiology* **59**, 1087–1091.

Drucker·, H., Burges, C. J. C., Kaufman, L., Smola··, A. & Vapoik, V. (1996), 'Support vector regression machines', *Advances in neural information processing systems* .

Du, J., Hu, M. & Zhang, W. (2020), 'Missing data problem in the monitoring system: A review', *IEEE Sensors Journal* **20**, 13984–13998.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. & Tabona, O. (2021), 'A survey on missing data in machine learning', *Journal of Big Data* **8**, 140.

Enders, C. K. (2022), *Applied Missing Data Analysis, Second Edition.*, 2nd ed. edn, Guilford Publications.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D. & Knudtson, M. L. (2002), 'Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses', *Journal of Clinical Epidemiology* **55**, 184–191.

Fasano, A., Rebaudo, G., Durante, D. & Petrone, S. (2019), 'A closed-form filter for binary time series', *arXiv* pp. 1–27.

Feng, H., Guoshun, C., Cheng, Y., Bingru, Y. & Yumei, C. (2005), 'A svm regression based approach to filling in missing values', *LNAI* **3683**, 581–587.

Feng, Z., Kobayashi, K. & Ainsworth, E. A. (2008), 'Impact of elevated ozone concentration on growth, physiology, and yield of wheat (¡i¿triticum aestivum¡/i¿ l.): a meta-analysis', *Global Change Biology* **14**, 2696–2708.

Folguera, L., Zupan, J., Cicerone, D. & Magallanes, J. F. (2015), 'Self-organizing maps for imputation of missing data in incomplete data matrices'.
URL: *http://dx.doi.org/10.1016/j.chemolab.2015.03.002*

Forouzanfar, M. H., Afshin, A., Alexander, L. T., Biryukov, S., Brauer, M., Cercy, K., Charlson, F. J., Cohen, A. J., Dandona, L., Estep, K., Ferrari, A. J., Frostad, J. J., Fullman, N., Godwin, W. W., Griswold, M., Hay, S. I., Kyu, H. H., Larson, H. J., Lim, S. S., Liu, P. Y., Lopez, A. D., Lozano, R., Marczak, L., Mokdad, A. H., Moradi-Lakeh, M., Naghavi, M., Reitsma, M. B., Roth, G. A., Sur, P. J., Vos, T., Wagner, J. A., Wang, H., Zhao, Y., Zhou, M., Barber, R. M., Bell, B., Blore, J. D., Casey, D. C., Coates, M. M., Cooperrider, K., Cornaby, L.,

Dicker, D., Erskine, H. E., Fleming, T., Foreman, K., Gakidou, E., Haagsma, J. A., Johnson, C. O., Kemmer, L., Ku, T., Leung, J., Masiye, F., Millear, A., Mirarefin, M., Misganaw, A., Mullany, E., Mumford, J. E., Ng, M., Olsen, H., Rao, P., Reinig, N., Roman, Y., Sandar, L., Santomauro, D. F., Slepak, E. L., Sorensen, R. J., Thomas, B. A., Vollset, S. E., Whiteford, H. A., Zipkin, B., Murray, C. J., Mock, C. N., Anderson, B. O., Futran, N. D., Anderson, H. R., Bhutta, Z. A., Nisar, M. I., Akseer, N., Krueger, H., Gotay, C. C., Kissoon, N., Kopec, J. A., Pourmalek, F., Burnett, R., Abajobir, A. A., Knibbs, L. D., Veerman, J. L., Lalloo, R., Scott, J. G., Alam, N. K., Gouda, H. N., Guo, Y., McGrath, J. J., Jeemon, P., Dandona, R., Goenka, S., Kumar, G. A., Gething, P. W., Bisanzio, D., Deribew, A., Darby, S. C., Ali, R., Bennett, D. A., Jha, V., Kinfu, Y., McKee, M., Murthy, G. V., Pearce, N., Stöckl, H., Duan, L., Jin, Y., Li, Y., Liu, S., Wang, L., Ye, P., Liang, X., Azzopardi, P., Patton, G. C., Meretoja, A., Alam, K., Borschmann, R., Colquhoun, S. M., Weintraub, R. G., Szoeke, C. E., Ademi, Z., Taylor, H. R., Wijeratne, T., Batis, C., Barquera, S., Campos-Nonato, I. R., Contreras, A. G., Cuevas-Nasu, L., De, V., Gomez-Dantes, H., Heredia-Pi, I. B., Medina, C., Mejia-Rodriguez, F., Hernandez, J. C. M., Razo-García, C. A., Rivera, J. A., Rodríguez-Ramírez, S., Sánchez-Pimienta, T. G., Servan-Mori, E. E., Shamah, T., Mensah, G. A., Hoff, H. J., Neal, B., Driscoll, T. R., Kemp, A. H., Leigh, J., Mekonnen, A. B., Bhatt, S., Fürst, T., Piel, F. B., Rodriguez, A., Hutchings, S. J., Majeed, A., Soljak, M., Salomon, J. A., Thorne-Lyman, A. L., Ajala, O. N., Bärnighausen, T., Cahill, L. E., Ding, E. L., Farvid, M. S., Khatibzadeh, S., Wagner, G. R., Shrime, M. G., Fitchett, J. R., Aasvang, G. M., Savic, M., Abate, K. H., Gebrehiwot, T. T., Gebremedhin, A. T., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abdulle, A. M., Abera, S. F., Melaku, Y. A., Abyu, G. Y., Betsu, B. D., Hailu, G. B., Tekle, D. Y., Yalew, A. Z., Abraham, B., Abu-Raddad, L. J., Adebiyi, A. O., Adedeji, I. A., Adou, A. K., Adsuar, J. C., Agardh, E. E., Rehm, J., Badawi, A., Popova, S., Agarwal, A., Ahmad, A., Akinyemiju, T. F., Schwebel, D. C., Singh, J. A., Al-Aly, Z., Aldhahri, S. F.,

Altirkawi, K. A., Terkawi, A. S., Aldridge, R. W., Tillmann, T., Alemu, Z. A., Tegegne, T. K., Alkerwi, A., Alla, F., Guillemin, F., Allebeck, P., Rabiee, R. H., Fereshtehnejad, S. M., Kivipelto, M., Carrero, J. J., Weiderpass, E., Havmoeller, R., Sindi, S., Alsharif, U., Alvarez, E., Alvis-Guzman, N., Amare, A. T., Ciobanu, L. G., Taye, B. W., Amberbir, A., Amegah, A. K., Amini, H., Karema, C. K., Ammar, W., Harb, H. L., Amrock, S. M., Andersen, H. H., Antonio, C. A., Faraon, E. J., Anwari, P., Ärnlöv, J., Larsson, A., Artaman, A., Asayesh, H., Asghar, R. J., Assadi, R., Atique, S., Avokpaho, E. F., Awasthi, A., Ayala, B. P., Bacha, U., Bahit, M. C., Balakrishnan, K., Barac, A., Barker-Collo, S. L., del Pozo-Cruz, B., Mohammed, S., Barregard, L., Petzold, M., Barrero, L. H., Basu, S., Del, L. C., Bazargan-Hejazi, S., Beardsley, J., Bedi, N., Beghi, E., Sheth, K. N., Bell, M. L., Huang, J. J., Bello, A. K., Santos, I. S., Bensenor, I. M., Lotufo, P. A., Berhane, A., Wolfe, C. D., Bernabé, E., Roba, H. S., Beyene, A. S., Hassen, T. A., Mesfin, Y. M., Bhala, N., Bhansali, A., Biadgilign, S., Bikbov, B., Bjertness, E., Htet, A. S., Boufous, S., Degenhardt, L., Resnikoff, S., Calabria, B., Bourne, R. R., Brainin, M., Brazinova, A., Majdan, M., Shen, J., Breitborde, N. J., Brenner, H., Schöttker, B., Broday, D. M., Brugha, T. S., Brunekreef, B., Kromhout, H., Butt, Z. A., van Donkelaar, A., Martin, R. V., Cárdenas, R., Carpenter, D. O., Castañeda-Orjuela, C. A., Castillo, J., Castro, R. E., Catalá-López, F., Chang, J., Chiang, P. P., Chibalabala, M., Chimed-Ochir, O., Jiang, Y., Takahashi, K., Chisumpa, V. H., Mapoma, C. C., Chitheer, A. A., Choi, J. J., Christensen, H., Christopher, D. J., Cooper, L. T., Crump, J. A., Poulton, R. G., Damasceno, A., Dargan, P. I., das Neves, J., Davis, A. C., Newton, J. N., Steel, N., Davletov, K., de Castro, E. F., De, D., Dellavalle, R. P., Des, D. C., Dharmaratne, S. D., Dhillon, P. K., Lal, D. K., Zodpey, S., Diaz-Torné, C., Dorsey, E. R., Doyle, K. E., Dubey, M., Rahman, M. H., Ram, U., Singh, A., Yadav, A. K., Duncan, B. B., Kieling, C., Schmidt, M. I., Elyazar, I., Endries, A. Y., Ermakov, S. P., Eshrati, B., Farzadfar, F., Kasaeian, A., Parsaeian, M., Esteghamati, A., Hafezi-Nejad, N., Sheikhbahaei, S., Fahimi, S., Malekzadeh, R., Roshandel, G., Sepanlou,

S. G., Hassanvand, M. S., Heydarpour, P., Rahimi-Movaghar, V., Yaseri, M., Farid, T. A., Khan, A. R., Farinha, C. S., Faro, A., Feigin, V. L., Fernandes, J. G., Fischer, F., Foigt, N., Shiue, I., Fowkes, F. G., Franklin, R. C., Garcia-Basteiro, A. L., Geleijnse, J. M., Jibat, T., Gessner, B. D., Tefera, W., Giref, A. Z., Haile, D., Manamo, W. A., Giroud, M., Gishu, M. D., Martinez-Raga, J., Gomez-Cabrera, M. C., Gona, P., Goodridge, A., Gopalani, S. V., Goto, A., Inoue, M., Gugnani, H. C., Gupta, R., Gutiérrez, R. A., Orozco, R., Halasa, Y. A., Undurraga, E. A., Hamadeh, R. R., Hamidi, S., Handal, A. J., Hankey, G. J., Hao, Y., Harikrishnan, S., Haro, J. M., Hernández-Llanes, N. F., Hoek, H. W., Tura, A. K., Horino, M., Horita, N., Hosgood, H. D., Hoy, D. G., Hsairi, M., Hu, G., Husseini, A., Huybrechts, I., Iburg, K. M., Idrisov, B. T., Kwan, G. F., Ileanu, B. V., Pana, A., Kawakami, N., Shibuya, K., Jacobs, T. A., Jacobsen, K. H., Jahanmehr, N., Jakovljevic, M. B., Jansen, H. A., Jassal, S. K., Stein, M. B., Javanbakht, M., Jayaraman, S. P., Jayatilleke, A. U., Jee, S. H., Jonas, J. B., Kabir, Z., Kalkonde, Y., Kamal, R., She, J., Kan, H., Karch, A., Karimkhani, C., Kaul, A., Kazi, D. S., Keiyoro, P. N., Parry, C. D., Kengne, A. P., Matzopoulos, R., Wiysonge, C. S., Stein, D. J., Mayosi, B. M., Keren, A., Khader, Y. S., Khan, E. A., Khan, G., Khang, Y. H., Won, S., Khera, S., Tavakkoli, M., Khoja, T. A., Khubchandani, J., Kim, C., Kim, D., Kimokoti, R. W., Kokubo, Y., Koul, P. A., Koyanagi, A., Kravchenko, M., Varakin, Y. Y., Kuate, B., Kuchenbecker, R. S., Kucuk, B., Kuipers, E. J., Lallukka, T., Shiri, R., Meretoja, T. J., Lan, Q., Latif, A. A., Lawrynowicz, A. E., Leasher, J. L., Levi, M., Li, X., Liang, J., Lloyd, B. K., Logroscino, G., Lunevicius, R., Pope, D., Mahdavi, M., Malta, D. C., Marcenes, W., Matsushita, K., Nachega, J. B., Tran, B. X., Meaney, P. A., Mehari, A., Tedla, B. A., Memish, Z. A., Mendoza, W., Mensink, G. B., Mhimbira, F. A., Miller, T. R., Mills, E. J., Mohammadi, A., Mola, G. L., Monasta, L., Morawska, L., Norman, R. E., Mori, R., Mozaff, D., Shi, P., Werdecker, A., Mueller, U. O., Paternina, A. J., Westerman, R., Seedat, S., Naheed, A., Nangia, V., Nassiri, N., Nguyen, Q. L., Nkamedjie, P. M., Norheim, O. F., Norrving, B., Nyakarahuka, L.,

Obermeyer, C. M., Ogbo, F. A., Oh, I., Oladimeji, O., Sartorius, B., Olusanya, B. O., Olivares, P. R., Olusanya, J. O., Opio, J. N., Oren, E., Ortiz, A., Ota, E., Mahesh, P. A., Park, E., Patel, T., Patil, S. T., Patten, S. B., Wang, J., Pereira, D. M., Cortinovis, M., Giussani, G., Perico, N., Remuzzi, G., Pesudovs, K., Phillips, M. R., Pillay, J. D., Plass, D., Tobollik, M., Polinder, S., Pond, C. D., Pope, C. A., Prasad, N. M., Qorbani, M., Radfar, A., Rafay, A., Rana, S. M., Rahman, M., Rahman, S. U., Rajsic, S., Rai, R. K., Raju, M., Ranganathan, K., Refaat, A. H., Rehm, C. D., Ribeiro, A. L., Rojas-Rueda, D., Roy, A., Satpathy, M., Tandon, N., Rothenbacher, D., Saleh, M. M., Sanabria, J. R., Sanchez-Riera, L., Sanchez-Niño, M. D., Sarmiento-Suarez, R., Sawhney, M., Schmidhuber, J., Schneider, I. J., Schutte, A. E., Silva, D. A., Shahraz, S., Shin, M., Shaheen, A., Shaikh, M. A., Sharma, R., Shigematsu, M., Yoon, S., Shishani, K., Sigfusdottir, I. D., Singh, P. K., Silveira, D. G., Silverberg, J. I., Yano, Y., Soneji, S., Stranges, S., Steckling, N., Sreeramareddy, C. T., Stathopoulou, V., Stroumpoulis, K., Sunguya, B. F., Swaminathan, S., Sykes, B. L., Tabarés-Seisdedos, R., Talongwa, R. T., Tanne, D., Tuzcu, E. M., Thakur, J., Shaddick, G., Thomas, M. L., Thrift, A. G., Thurston, G. D., Thomson, A. J., Topor-Madry, R., Topouzis, F., Towbin, J. A., Uthman, O. A., Tobe-Gai, R., Tsilimparis, N., Tsala, Z., Tyrovolas, S., Ukwaja, K. N., van Os, J., Vasankari, T., Venketasubramanian, N., Violante, F. S., Waller, S. G., Uneke, C. J., Wang, Y., Weichenthal, S., Woolf, A. D., Xavier, D., Xu, G., Yakob, B., Yip, P., Kesavachandran, C. N., Montico, M., Ronfani, L., Yu, C., Zaidi, Z., Yonemoto, N., Younis, M. Z., Wubshet, M., Zuhlke, L. J., Zaki, M. E. & Zhu, J. (2016), 'Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the global burden of disease study 2015', *The Lancet* **388**, 1659–1724.

Geweke, J. (1989), 'Bayesian inference in econometric models using monte carlo integration', *Econometrica: Journal of the Econometric Society* pp. 1317–1339.

Ghahramani, Z. & Jordan, M. I. (1995), 'Supervised learning from incomplete data via an em approach', *Proc. Adv. Neural Inf. Process. Syst.* **6**, 120–127.

Gimpy, Vohra, D. R. & Minakshi (2014), 'Estimation of missing values using decision tree approach', *Int J Comput Sci Inf Technol* **5**, 5216–5220.
URL: *www.ijcsit.com*

Gnauck, A. & Luther, B. (2005), 'Missing data in environmental time series-a problem analysis'.

González-Farías, G., Domínguez-Molina, A. & Gupta, A. K. (2004), 'Additive properties of skew normal random vectors', *Journal of Statistical Planning and Inference* **126**, 521–534.

Guhaniyogi, R., Li, C., Savitsky, T. & Srivastava, S. (2019), 'A divide-and-conquer bayesian approach to large-scale kriging'.

Guljanov, G., Mutschler, W. & Trede, M. (2022), 'Dynare working papers series pruned skewed kalman filter and smoother: With application to the yield curve pruned skewed kalman filter and smoother: With application to the yield curve'.
URL: *https://www.cepremap.fr*

Hartley, H. O. & Hocking, R. R. (1971), 'The analysis of incomplete data', *Biometrics* **27**, 783.

Heemink, A. W. & Segers, A. J. (2002), 'Modeling and prediction of environmental data in space and time using kalman filtering', *Stochastic Environmental Research and Risk Assessment* **16**, 225–240.

Hensman, J., Fusi, N. O. & Lawrence, N. D. (2013), 'Gaussian processes for big data'.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla,

S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S. & Thépaut, J. (2020), 'The era5 global reanalysis', *Quarterly Journal of the Royal Meteorological Society* **146**, 1999–2049.

Hoffman, M. D. & Gelman, A. (2014), 'The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo', *Journal of Machine Learning Research* **15**, 1593–1623.
URL: *http://mcmc-jags.sourceforge.net*

Hoi, K. I., Yuen, K. V. & Mok, K. M. (2008), 'Kalman filter based prediction system for wintertime pm10 concentrations in macau', *Global Nest Journal* **10**, 140–150.

Huang, J. & Sun, H. (2016), Grey relational analysis based k nearest neighbor missing data imputation for software quality datasets, IEEE, pp. 86–91.
URL: *http://ieeexplore.ieee.org/document/7589788/*

Huangfu, P. & Atkinson, R. (2020), 'Long-term exposure to no2 and o3 and all-cause and respiratory mortality: A systematic review and meta-analysis', *Environment International* **144**, 105998.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M. & Suttie, M. (2019), 'The cams reanalysis of atmospheric composition', *Atmospheric Chemistry and Physics* **19**, 3515–3556.

Inness, A., Blechschmidt, A.-M., Bouarar, I., Chabrillat, S., Crepulja, M., Engelen, R. J., Eskes, H., Flemming, J., Gaudel, A., Hendrick, F., Huijnen, V., Jones, L., Kapsomenakis, J., Katragkou, E., Keppens, A., Langerock, B., de Mazière, M.,

Melas, D., Parrington, M., Peuch, V. H., Razinger, M., Richter, A., Schultz, M. G., Suttie, M., Thouret, V., Vrekoussis, M., Wagner, A. & Zerefos, C. (2015), 'Data assimilation of satellite-retrieved ozone, carbon monoxide and nitrogen dioxide with ecmwf's composition-ifs', *Atmospheric Chemistry and Physics* **15**, 5275–5303.

IPCC (2014), *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.*
URL: *www.cambridge.org*

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. & Kolehmainen, M. (2004), 'Methods for imputation of missing values in air quality data sets', *Atmospheric Environment* **38**, 2895–2907.

Kalman, R. E. (1960), 'A new approach to linear filtering and prediction problems', *Journal of Fluids Engineering, Transactions of the ASME* **82**, 35–45.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M. & Potter, G. L. (2002), 'Ncep–doe amip-ii reanalysis (r-2)', *Bulletin of the American Meteorological Society* **83**, 1631–1644.

Kiani, K. & Saleem, K. (2017), 'K-nearest temperature trends: A method for weather temperature data imputation'.
URL: *http://dx.doi.org/10.1145/3077584.3077592*

Lamrini, B., Lakhal, E.-K., Lann, M.-V. L. & Wehenkel, L. (2011), 'Data validation and missing data reconstruction using self-organizing map for water treatment', *Neural Computing and Applications* **20**, 575–588.
URL: *http://link.springer.com/10.1007/s00521-011-0526-5*

Lao, J. & Louf, R. (2020), 'Blackjax: A sampling library for jax'.
URL: *http://github.com/blackjax-devs/blackjax*

Lee, J.-W. & Park, S.-C. (2015), 'Artificial neural network-based data recovery system for the time series of tide stations', *Journal of Coastal Research* **32**, 213.
URL: *https://bioone.org/journals/journal-of-coastal-research/volume-32/issue-1/JCOASTRES-D-14-00233.1/Artificial-Neural-Network-Based-Data-Recovery-System-for-the-Time/10.2112/JCOASTRES-D-14-00233.1.full*

Little, R. J. A. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, John Wiley and Sons, Inc.

Liu, B., Yang, J., Yuan, J., Wang, J., Dai, Q., Li, T., Bi, X., Feng, Y., Xiao, Z., Zhang, Y. & Xu, H. (2017), 'Source apportionment of atmospheric pollutants based on the online data by using pmf and me2 models at a megacity, china', *Atmospheric Research* **185**, 22–31.
URL: *http://dx.doi.org/10.1016/j.atmosres.2016.10.023*

Liu, J., Xie, Q., Liu, G. & Sun, Y. (2016), 'A method for missing data recovery of air pollutants monitoring in henhouse based on qgsa-svm information monitoring and processing in agriculture view project a method for missing data recovery of air pollutants monitoring in henhouse based on qgsa-svm', *International Journal of Smart Home* **10**, 139–148.
URL: *http://dx.*

Liu, J., Xie, Q. & Zhang, Y. (2015), 'A method for missing data recovery of waste gas monitoring in animal building based on ga-svm', *International Journal of Smart Home* **9**, 175–184.
URL: *http://dx.doi.org/10.14257/ijsh.2015.9.5.17*

Lovett, G., Burns, D., Driscoll, C., Jenkins, J., Mitchell, M., Rustad, L., Shanley, J., Likens, G. & Haeuber, R. (2007), 'Who needs environmental monitoring?', *Frontiers in Ecology and the Environment* pp. 253–260.

Mallasto, A. & Feragen, A. (2017), 'Learning from uncertain curves: The 2-wasserstein metric for gaussian processes', *NIPS* .

Meijering, E. (2002), 'A chronology of interpolation: from ancient astronomy to modern signal and image processing', *Proceedings of the IEEE* **90**, 319–342.
URL: *http://ieeexplore.ieee.org/document/993400/*

Meng, X.-L. & Rubin, D. B. (1992), 'Performing likelihood ratio tests with multiply-imputed data sets', *Biometrika* **79**, 103–114.
URL: *https://academic.oup.com/biomet/article/79/1/103/285615*

Metia, S., Oduro, S. D. & Sinha, A. P. (2020), Pollutant profile estimation using unscented kalman filter, Vol. 591, Springer, pp. 17–28.

Mirzaei, A., Carter, S. R., Patanwala, A. E. & Schneider, C. R. (2022), 'Missing data in surveys: Key concepts, approaches, and applications', *Research in Social and Administrative Pharmacy* **18**, 2308–2316.

Napelenok, S. L., Pinder, R. W., Gilliland, A. B. & Martin, R. V. (2008), 'A method for evaluating spatially-resolved nox emissions using kalman filter inversion, direct sensitivities, and space-based no2 observations', *Atmospheric Chemistry and Physics* **8**, 5603–5614.

Naveau, P., Genton, M. G. & Shen, X. (2005), 'A skewed kalman filter', *Journal of Multivariate Analysis* **94**, 382–400.

NCEH (2020), 'Preparing for the regional health impacts of climate change in the united states'.
URL: *https://www.cdc.gov/climateandhealth/*

Nkiaka, E., Nawaz, N. R. & Lovett, J. C. (2016), 'Using self-organizing maps to infill missing data in hydro-meteorological time series from the logone catchment, lake chad basin', *Environmental Monitoring and Assessment* **188**, 400.
URL: *http://link.springer.com/10.1007/s10661-016-5385-1*

Noor, N. M., Abdullah, M. M. A. B., Yahaya, A. S. & Ramli, N. A. (2014), 'Comparison of linear interpolation method and mean method to replace the

missing values in environmental data set', *Materials Science Forum* **803**, 278–281.

Oehmcke, S., Zielinski, O. & Kramer, O. (2016), knn ensembles with penalized dtw for multivariate time series imputation, IEEE, pp. 2774–2781.

Ohba, K., Yoneda, Y., Kurihara, K., Suganuma, T., Ito, H., Ishihara, N., Gotoh, K., Yamashita, K. & Masu, K. (2016), 'Polynomial regression techniques for environmental data recovery in wireless sensor networks', *Sensors and Transducers* **199**, 1–9.
URL: *http://www.sensorsportal.com*

Organization, W. H. (2022), 'Ambient (outdoor) air pollution [fact sheet]'.
URL: *https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health*

Otero, N., Sillmann, J., Schnell, J. L., Rust, H. W. & Butler, T. (2016), 'Letter • open access synoptic and meteorological drivers of extreme ozone concentrations over europe you may also like climate change penalty and benefit on surface ozone: a global perspective based on cmip6 earth system models synoptic and meteorological drivers of extreme ozone concentrations over europe', *Environ. Res. Lett* **11**, 24005.

Pan, L. & Li, J. (2010), 'K-nearest neighbor based missing data estimation algorithm in wireless sensor networks', *Wireless Sensor Network* **02**, 115–122.
URL: *http://www.scirp.org/journal/PaperInformation.aspx?PaperID=1378*

Petris, G., Petrone, S. & Campagnoli, P. (2009), 'Dynamic linear models'.

Pinder, T. & Dodd, D. (2022), 'Gpjax: A gaussian process framework in jax', *Journal of Open Source Software* **7**, 4455.
URL: *https://joss.theoj.org/papers/10.21105/joss.04455*

Plaia, A. & Bondì, A. L. (2006), 'Single imputation method of missing values in environmental pollution data sets', *Atmospheric Environment* **40**, 7316–7330.

Polikar, R. (2006), 'Ensemble based systems in decision making; ensemble based systems in decision making'.

Pruss-Ustun, A., Corvalan, C. C., Bos, R., Neira, M. & Organization, W. H. (2018), 'Preventing disease through healthy environments : a global assessment of the burden of disease from environmental risks', p. 147.
URL: *https://www.who.int/publications/i/item/9789241565196*

Qu, L., Li, L., Zhang, Y. & Hu, J. (2009), 'Ppca-based missing data imputation for traffic flow volume: A systematical approach', *IEEE Transactions on intelligent transportation systems* **10**, 512–522.

Quiñonero-Candela, J., Ramussen, C. E. & Williams, C. K. I. (2007), 'Approximation methods for gaussian process regression'.

Quiñonero, J., Quiñonero-Candela, Q., Rasmussen, C. E. & De, C. M. (2005), 'A unifying view of sparse approximate gaussian process regression', *Journal of Machine Learning Research* **6**, 1939–1959.

Rahman, M. G. & Islam, M. Z. (2011), 'A decision tree-based missing value imputation technique for data pre-processing'.

Rahman, M. G. & Islam, M. Z. (2013), 'Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques', *Knowledge-Based Systems* **53**, 51–65.

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, The MIT Press.
URL: *www.GaussianProcess.org/gpml*

Ridder, K. D., Kumar, U., Lauwaet, D., Blyth, L. & Lefebvre, W. (2012), 'Kalman filter-based air quality forecast adjustment', *Atmospheric Environment* **50**, 381–384.
URL: *http://dx.doi.org/10.1016/j.atmosenv.2012.01.032*

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–92.
URL: *https://academic.oup.com/biomet/article/63/3/581/270932*

Rubin, D. B. (1996), 'Multiple imputation after 18+ years', *Journal of the American statistical Association* **91**, 473–489.

Rubin, D. B. (2004), *Multiple imputation for nonresponse in surveys*, John Wiley & Sons.

Rue, H. & Held, L. (2005), *Gaussian Markov random fields: theory and applications*, CRC press.

Schafer, J. L. (1997), *Analysis of incomplete multivariate data*, CRC press.

Schafer, J. L. & Graham, J. W. (2002), 'Missing data: Our view of the state of the art'.

Schneider, T. (2001), 'Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values', *Journal of Climate* **14**, 853–871.

Schoellhamer, D. H. (2001), 'Singular spectrum analysis for time series with missing data', *Geophysical Research Letters* **28**, 3187–3190.
URL: *https://onlinelibrary.wiley.com/doi/full/10.1029/2000GL012698*
https://onlinelibrary.wiley.com/doi/abs/10.1029/2000GL012698
https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2000GL012698

Schultz, M. G., Schröder, S., Lyapina, O., Cooper, O. R., Galbally, I., Petropavlovskikh, I., von Schneidemesser, E., Tanimoto, H., Elshorbany, Y.,

Naja, M., Seguel, R. J., Dauert, U., Eckhardt, P., Feigenspan, S., Fiebig, M., Hjellbrekke, A.-G., Hong, Y.-D., Kjeld, P. C., Koide, H., Lear, G., Tarasick, D., Ueno, M., Wallasch, M., Baumgardner, D., Chuang, M.-T., Gillett, R., Lee, M., Molloy, S., Moolla, R., Wang, T., Sharps, K., Adame, J. A., Ancellet, G., Apadula, F., Artaxo, P., Barlasina, M. E., Bogucka, M., Bonasoni, P., Chang, L., Colomb, A., Cuevas-Agulló, E., Cupeiro, M., Degorska, A., Ding, A., Fröhlich, M., Frolova, M., Gadhavi, H., Gheusi, F., Gilge, S., Gonzalez, M. Y., Gros, V., Hamad, S. H., Helmig, D., Henriques, D., Hermansen, O., Holla, R., Hueber, J., Im, U., Jaffe, D. A., Komala, N., Kubistin, D., Lam, K.-S., Laurila, T., Lee, H., Levy, I., Mazzoleni, C., Mazzoleni, L. R., McClure-Begley, A., Mohamad, M., Murovec, M., Navarro-Comas, M., Nicodim, F., Parrish, D., Read, K. A., Reid, N., Ries, L., Saxena, P., Schwab, J. J., Scorgie, Y., Senik, I., Simmonds, P., Sinha, V., Skorokhod, A. I., Spain, G., Spangl, W., Spoor, R., Springston, S. R., Steer, K., Steinbacher, M., Suharguniyawan, E., Torre, P., Trickl, T., Weili, L., Weller, R., Xiaobin, X., Xue, L. & Zhiqiang, M. (2017), 'Tropospheric ozone assessment report: Database and metrics data of global surface ozone observations', *Elementa: Science of the Anthropocene* **5**.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I. & McCulloch, R. E. (2016), 'Bayes and big data: The consensus monte carlo algorithm', *International Journal of Management Science and Engineering Management* **11**, 78–88.

Shang, Q., Yang, Z., Gao, S. & Tan, D. (2018), 'An imputation method for missing traffic data based on fcm optimized by pso-svr'.
URL: *https://doi.org/10.1155/2018/2935248*

Shen, Y., Peng, F. & Li, B. (2015), 'Improved singular spectrum analysis for time series with missing data', *Nonlin. Processes Geophys* **22**, 371–376.
URL: *www.nonlin-processes-geophys.net/22/371/2015/*

Särkkä, S. (2013), *Bayesian filtering and smoothing*, Cambridge University Press.

Tarasick, D., Galbally, I. E., Cooper, O. R., Schultz, M. G., Ancellet, G., Leblanc, T., Wallington, T. J., Ziemke, J., Liu, X., Steinbacher, M., Staehelin, J., Vigouroux, C., Hannigan, J. W., García, O., Foret, G., Zanis, P., Weatherhead, E., Petropavlovskikh, I., Worden, H., Osman, M., Liu, J., Chang, K.-L., Gaudel, A., Lin, M., Granados-Muñoz, M., Thompson, A. M., Oltmans, S. J., Cuesta, J., Dufour, G., Thouret, V., Hassler, B., Trickl, T. & Neu, J. L. (2019), 'Tropospheric ozone assessment report: Tropospheric ozone from 1877 to 2016, observed levels, trends and uncertainties', *Elementa: Science of the Anthropocene* **7**.

Titsias, M. K. (2009), 'Variational learning of inducing variables in sparse gaussian processes'.

Twala, B. (2009), 'An empirical comparison of techniques for handling incomplete data using decision trees', *Applied Artificial Intelligence* **23**, 373–405.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., İlhan Polat, Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee,

P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y. O. & Vázquez-Baeza, Y. (2020), 'Scipy 1.0: fundamental algorithms for scientific computing in python', *Nature Methods* **17**, 261–272.

Vyner, C., Nemeth, C. & Sherlock, C. (2022), 'Swiss: A scalable markov chain monte carlo divide-and-conquer strategy a preprint'.

Wagner, A., Bennouna, Y., Blechschmidt, A.-M., Brasseur, G., Chabrillat, S., Christophe, Y., Errera, Q., Eskes, H., Flemming, J., Hansen, K. M., Inness, A., Kapsomenakis, J., Langerock, B., Richter, A., Sudarchikova, N., Thouret, V. & Zerefos, C. (2020), 'Comprehensive evaluation of the copernicus atmosphere monitoring service (cams) reanalysis against independent observations: Reactive gases'.
URL: *https://doi.org/10.1525/elementa.2020.00171*

Wang, K., Arellano-Valle, R. B., Azzalini, A. & Genton, M. G. (2023), 'On the non-identifiability of unified skew-normal distributions'.

Wittig, V. E., Ainsworth, E. A., Naidu, S. L., Karnosky, D. F. & Long, S. P. (2009), 'Quantifying the impact of current and future tropospheric ozone on tree biomass, growth, physiology and biochemistry: a quantitative meta-analysis', *Global Change Biology* **15**, 396–424.
URL: *https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2486.2008.01774.x*

Wu, C. & Robert, C. (2017), 'Average of recentered parallel mcmc for big data'.

Zhang, K., Gonzalez, R., Huang, B. & Ji, G. (2015), 'Expectation–maximization approach to fault diagnosis with missing data', *IEEE Transactions on Industrial Electronics* **62**, 1231–1240.
URL: *http://ieeexplore.ieee.org/document/6850032/*

Zhang, X.-F., Ou-Yang, L., Yang, S., Zhao, X.-M., Hu, X. & Yan, H. (2019), 'Enimpute: imputing dropout events in single-cell rna-sequencing data via ensemble learning', *Bioinformatics* **35**, 4827–4829.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C. & Baklanov, A. (2012), 'Real-time air quality forecasting, part i: History, techniques, and current status', *Atmospheric Environment* **60**, 632–655.
URL: *http://dx.doi.org/10.1016/j.atmosenv.2012.06.031*

Zhang, Y. & Thorburn, P. J. (2022), 'Handling missing data in near real-time environmental monitoring: A system and a review of selected methods', *Future Generation Computer Systems* **128**, 63–72.

Zolghadri, A. & Cazaurang, F. (2006), 'Adaptive nonlinear state-space modelling for the prediction of daily mean pm10 concentrations', *Environmental Modelling and Software* **21**, 885–894.