A thesis submitted in partial fulfilment of the requirements for the degree of

Master in Econometrics & Management Science

Linear Mixed-Effect State-Space Forecasting: Completing the algorithm

Nathan Schot

(457066)

Supervisor: dr. M. van der Wel

Second Assesor: dr. J.W.N. Reuvers

October 8, 2018

Abstract

Statistical downscaling techniques are used to translate global climate scenarios into local impact forecasts. In this study, I propose a new algorithm for downscaling by using a linear mixed-effect state-space model (LMESS). The rationale to use this model in a climate data context is that it allows for both time-varying and fixed relations between dependent and explanatory variables. My findings show the importance of identifying the correct random and fixed effects. I develop a new method for selection based on the state-space formulation with fixed parameters by Chow (1984). I apply the proposed methods to climate data at five different weather stations in the Netherlands. My findings show that the LMESS model is not able to consistently outperform a multivariate linear regression forecast method.

Keywords: Linear Mixed-effect, State-Space, variable selection, forecasting, climate

Contents

1 Introduction

Improving the ability to predict weather events becomes increasingly important, as the coastal regions of Europe, and especially the Netherlands, are highly susceptible to extreme weather events (Beniston et al., 2007). In this region, the occurrence of hurricanes (Haarsma et al., 2013) and extreme heat waves (Beniston, 2004) is on the rise. These events cause both natural and economic damage and will majorly impact billions of people (Dorland, Tol, & Palutikof, 1999). Global Circulation Models (GCMs) are an important tool in the assessment of climate change, but have low resolution, making them unable to predict local impacts. A considerable amount of research has therefore gone into answering the question: "How do we scale down global climate models to make local impact forecasts?" The methods that provide an answer to this question are called downscaling methods.

Multiple downscaling techniques have been developed to bridge the gap between global forecasts and local impacts. Most research has gone into linear regression methods, as Fowler, Blenkinsop, & Tebaldi (2007) show. However, Kokic, Crimp, & Howden (2011) show that a linear mixed-effect state-space (LMESS) approach provides better predictions of rainfall and temperature than linear regression in Australia. The rationale behind their results is that the LMESS model generalises the linear regression model by allowing smooth time variation of regression coefficients. Furthermore, the LMESS model allows for forecasting of non-stationary time series, which is a known feature of climate variables (Tank, Zwiers, & Zhang, 2009). The LMESS model is also a generalisation of the linear state-space model, but has less chance of over-fitting, because it allows for a subset of parameters to remain fixed.

In this research, I build on the steps taken by Kokic et al. (2011) to propose an algorithm to forecast climate variables using the linear mixed-effect state-space model. The contributions made to the existing literature are threefold. Firstly, I investigate the remark made by Kokic et al. (2011) that the manual variable selection procedure used in their study is suboptimal and propose two variable selection algorithms never before applied in the context of linear mixed-effect state-space models. One is based on the state-space aspect of the LMESS model by adjusting the conventional linear state-space model and using the estimation framework developed by Chow (1984) and Durbin & Koopman (2012). From the mixed-effect aspect of the LMESS model stems the second variable selection method I evaluate, which is an adjusted form of the variable selection

algorithm proposed by Bondell, Krishna, & Ghosh (2010).

Secondly, I examine the influence of incorrect model selection on forecasting with the linear mixed-effect state-space model. The susceptibility of forecast accuracy to model specification has been studied in other contexts, such as auto-regressive conditional heteroskedasticity (Nelson & Foster, 1995), standard volatility (Andersen & Bollerslev, 1998), and neural networks (Swanson & White, 1997), but not yet for the LMESS model. Evaluating the forecasting performance of wrong model specifications relative to the true model highlights the necessity of an accurate variable selection algorithm.

Thirdly, I study the application of the proposed linear mixed-effect state-space forecast algorithm to a new selection of observations. In this research, I use data from five weather stations in the Netherlands. This extends the research of Kokic et al. (2011) by studying a different climate than the Australian climate considered in their study. For this reason, I use a new selection of explanatory covariates. Furthermore, I shorten the temporal distance between measurements to one month to investigate the capability of the linear mixed-effect state-space model to forecast on a shorter time scale.

Research into climate change and its consequences is mainly focused on global or continental scale. Wetherald & Manabe (1995) show the influence of rising CO_2 levels on lack of precipitation in summer under a variety of circumstances using an idealised model for global geography. Another example is the research of Hoerling, Hurrell, & Xu (2001), who identify the North Atlantic Oscillation to be a driving factor of climate change across the North Atlantic region. A conclusion that was supported by later research of Cassou, Terray, & Phillips (2005), who studied the influence of the tropical region of the Atlantic ocean on climate regimes in Europe. Although these studies show a general long-term climate trend, the translation to local impacts is not addressed. Convery & Wagner (2015) argue that research on a local scale is at least as important, since improving forecasts and reducing uncertainty helps policy makers to develop appropriate measures to reduce climate risk. Local impact studies help answer questions such as: "Should the height of dikes be increased to account for more extreme water level fluctuations?", and "Does water reserve capacity need to be higher to be prepared for spells of draught?" From a food production perspective, accurate local impact forecasts are especially important, as operational and strategic decisions in the agricultural sector rely heavily on long range weather forecasts (Calanca, Bolius, Weigel, & Liniger, 2011). Downscaling techniques are

considered as the most promising method to bridge the gap between GCMs and local impacts. Another approach would be to extend the forecast horizon of weather prediction models, which is infeasible with the current methods used for forecasting. In their study, Kukkonen et al. (2012) compare different weather prediction methods used across Europe. These methods give forecasts based on analysis of the physical processes in the atmosphere and fall under the term chemical weather forecasting (CWF). The current models combine numerical weather prediction (NWP) and atmospheric chemistry simulations. Although the accuracy of these models in day-to-day forecasting is high, they are heavily reliant on high-resolution atmospheric observations. For example, the LOTOS-EUROS model, which is used in the Netherlands, uses 3D fields for wind direction, wind speed, temperature, and humidity (Schaap et al., 2008). Observation errors are therefore the main drawback of CWF techniques. Furthermore, Kukkonen et al. (2012) remark the inability of NWP models to incorporate all physical processes that determine weather changes, which makes accurate long-range forecasting difficult.

Beckmann & Buishand (2002) provide the sole research to date on the application of downscaling techniques in the Netherlands. Their paper shows the ability of a variety of regression models to forecast rainfall occurrence at five measurement sites across the Netherlands and Germany. However, Beckmann & Buishand (2002) remark that their modelling framework works best when rainfall occurrence and rainfall on wet days are analysed separately, which restricts the applicability of their forecasting methods. The work of Beckmann & Buishand (2002) is included in the overview of downscaling techniques comprised by Fowler et al. (2007). They summarise a wide variety of methods that have been investigated, from multivariate linear regression to neural networks, that aim to translate general circulation model predictions to local impacts. Across all research cited by Fowler et al. (2007), a wide variety of predictors is identified. However, the dynamic relation between dependent variables and predictors is a major hurdle for accurate modelling and forecasting.

With a simulation study, I show that my research provides another step to accurate forecasting with the linear mixed-effect state-space framework. I find that the variable selection procedure based on the state-space model formulation by Chow (1984) shows a model selection accuracy above 70%, where the method used by Kokic et al. (2011) finds the correct model in at most 20% of simulated scenarios. Furthermore, my proposed

algorithm is robust under different parameter values in the data generating process, with accuracy of at least 53% for a majority of different parameter scenarios. Adjusting the variable selection algorithm proposed by Bondell et al. (2010) to incorporate the state-space formulation does not consistently give the desired accuracy. The algorithm shows selection accuracy of up to 76%, but fails to correctly identify a single model specification across 100 simulated sets of data in a quarter of considered scenarios. Furthermore, my results show that the linear mixed-effect state-space model forecasts are sensitive to wrong model specifications. Across a variety of wrong model specifications, none can significantly outperform the true model specification in forecast accuracy. I thus show the necessity of an accurate variable selection algorithm when considering forecasting using the linear mixed-effect state-space model.

Application of my methods to climate data in the Netherlands shows that the linear mixed-effect state-space model can capture differences in time series dynamics between weather stations. Furthermore, forecasting with the LMESS model can reduce the root mean squared forecast error compared to a naive climatology forecast by up to 7% for the proportion of rainy days per month in De Bilt. On the other hand, in case of the mean rainfall, maximum temperature, and minimum temperature, the LMESS approach does not significantly improve climatology forecasts. For weather stations in De Kooy, Eelde, Vlissingen, and Beek, the linear mixed-effect state-space model does not improve on climatology forecasts significantly. Compared to a multivariate linear regression approach, the LMESS model does not significantly outperform the multivariate linear regression model in terms of forecast accuracy for any dependent variable at any of the five considered measurement sites.

The remainder of this thesis is organised as follows. In section 2 I introduce the linear mixed-effect state-space model and the estimation of parameters in that framework, as well as the variable selection methods examined in this study. I proceed in section 3 with a simulation study on variable selection accuracy and the influence of wrong model selection on forecast performance. Thereafter, I introduce the climate data used in this study in section 4. The application of the proposed linear mixed-effect state-space algorithm in the Netherlands is reported in section 5. Section 6 reports the conclusion of the research.

2 Methodology

This research aims to propose a linear mixed-effect state-space model (LMESS) estimation algorithm. Until now, no method has been proven to accurately identify significant explanatory covariates in case of the LMESS model. I propose such a method by extending methods used in either a linear state-space or a mixed-effect model context. In this section I introduce the formulation of the linear mixed-effect state-space model and the estimation of model parameters as derived by Kokic et al. (2011). Thereafter, I propose three distinct algorithms to identify fixed and random effects in the LMESS model. I end this section with forecasting equations of the linear mixed-effect state-space framework.

2.1 The Linear Mixed-Effect State-Space Model

The linear mixed-effect state-space model described by Kokic et al. (2011) is a state-space model where the observation equation is formulated as a linear mixed-effect model. In this research the mixed-effect terminology refers to the inclusion of both fixed and time-varying (random) coefficients in the observation equation. The state-space part of the model name comes from the state equation that describes the temporal relation between the time-varying coefficients. The representation of variables in a linear mixed-effect state-space form is a generalisation of two models commonly used in time series analysis, namely the multivariate linear regression model and the linear state-space model. Kokic et al. (2011) combine both specifications to formulate the LMESS model, which circumvents the main drawbacks of the individual models that limit applicability for climate forecasting.

Referring to time-varying model coefficients as random stems from the conventional linear mixed-effect model without state equation. Research using this formulation assumes there is a common fixed response coefficient for all research subjects. On the other hand, the linear mixed-effect model allows other response parameters to vary between different subjects. The between subject variation is assumed to follow a distribution from which all subject responses are randomly drawn, hence the name random effect. Throughout this research random and time-varying are used interchangeably.

Kokic et al. (2011) use the linear mixed-effect state-space model to forecast summary statistics of climate variable distributions by using a number of explanatory covariates. Just like the linear state-space model, the LMESS model is specified in terms of an observation

equation and a state equation. Suppose there are r climate variables observed at m different measurement sites. We have observed related scalar covariates $u_{0t}, u_{1t}, u_{2t}, \ldots$ over a time period $t \in \{1, \ldots, T\}$. The covariate $u_{0t} = 1$ is an intercept covariate per convention, included a priori to capture effects that the time-varying covariates can not capture. Then at each time t, the observation and state equation of the LMESS model for site $j \in \{1, \ldots, m\}$ are written as

$$y_{jt} = X_{jt}\beta_j + Z_{jt}\alpha_{jt} + v_{jt} \tag{1}$$

$$\alpha_{jt} = A_j\alpha_{jt-1} + w_{jt}, \tag{2}$$

where y_{jt} is a column vector containing the r observed dependent variables at time t for site j. The value corresponding to dependent variable i in this notation is y_{ijt}. The dependent variables may be pre-transformed to improve model fit. Matrices X_{jt} and Z_{jt} are block diagonal matrices containing covariates $u_{0t}, u_{1t}, u_{2t}, \ldots$ corresponding to the fixed and random effects, respectively. These matrices can be written in full as

$$X_{jt} = \begin{pmatrix} X_{1jt} & 0 & \cdots & 0 \\ 0 & X_{2jt} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & X_{rjt} \end{pmatrix}, \text{ and } Z_{jt} = \begin{pmatrix} Z_{1jt} & 0 & \cdots & 0 \\ 0 & Z_{2jt} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & Z_{rjt} \end{pmatrix}, \tag{3}$$

where X_{ijt} is a row vector of p^*_{ij} covariates included as fixed effects for observation y_{ijt}. The vector Z_{ijt} a row vector of q^*_{ij} covariates corresponding to the random effect on the same observation. The numbers p^*_{ij} and q^*_{ij} are the optimal number of fixed and random effect covariates for y_{ijt}, respectively. These are not known a priori, but are found from p_{ij} and q_{ij} initial candidate covariates out of the set of all covariates $u_{0t}, u_{1t}, u_{2t}, \ldots$. The vector α_{jt} has dimensions $(\sum_i q^*_{ij} \times 1)$ and contains the unobserved state parameters corresponding to the random effects. The matrix A_j is a $(\sum_i q^*_{ij} \times \sum_i q^*_{ij})$ state transition matrix. The vector $\beta_j = (\beta'_{1j}, \ldots, \beta'_{rj})'$ represents a $(\sum_i p^*_{ij} \times 1)$ fixed effect coefficient vector. The observation error v_{jt} is a $(r \times 1)$ vector which is normally distributed with mean zero and $(r \times r)$ covariance matrix R_j. The error of the state equation w_{jt} is a normally distributed $(\sum_i q^*_{ij} \times 1)$ vector with mean zero and $(\sum_i q^*_{ij} \times \sum_i q^*_{ij})$ covariance matrix Q_j. The model formulation does not impose any assumptions other than the normality of the errors. This means that parameter values and dynamics are allowed to vary between the m different measurement sites.

Let us consider an example to illustrate the formulation of the linear mixed-effect state-space model. Suppose there are $r = 4$ dependent variables measured at $j = De\,Bilt$, which result in observation vector y_{jt}. In this example, y_{1jt} corresponds to the number of rainy days in a month and y_{2jt} is the average rainfall on a rainy day. The variables y_{3jt} and y_{4jt} are the average maximum and minimum temperature per month, respectively. Besides the constant intercept $u_{0t} = 1$, there are four potential explanatory time-varying covariates $u_{1t}, \ldots, u_{4t}$ corresponding to atmospheric CO_2 level, the NAO index, cyclone density and anticyclone density, respectively. Suppose the model has the following relations: there are no random effect terms for y_{1jt}, but CO_2 level and the NAO index are fixed effects. The variable y_{2jt} has anticyclone density as random effect and a fixed intercept. The CO_2 level and the cyclone density are fixed effects for y_{3jt} and it has a random effect intercept. Lastly, y_{4jt} has the intercept and CO_2 level as random effect terms, and the NAO index and anticyclone density as fixed effects. Then the matrices X_{jt} and Z_{jt} are

$$X_{jt} = \begin{pmatrix} u_{1t} & u_{2t} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & u_{1t} & u_{3t} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & u_{2t} & u_{4t} \end{pmatrix}, \text{ and } Z_{jt} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ u_{4t} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & u_{1t} \end{pmatrix}.$$

Each covariate appears at most once in any row of X_{jt} and Z_{jt}, which is the most important assumption in the formulation of Kokic et al. (2011): Any covariate represents either a fixed or a random effect if it is included, so it can never be both fixed and random effect for a dependent variable.

The LMESS formulation poses two challenges. First, an algorithm is needed to find maximum likelihood estimates for all model parameters, as I need to account for the fixed effect term when using existing methods for linear state-space models. Second, to reduce the number of parameters to estimate, I need to find which explanatory covariates need to be treated as random effects, which are best considered fixed effects, and which may be excluded from the model.

2.2 Parameter estimation

In this section I outline the parameter estimation equations and an EM-algorithm that incorporates the fixed effects term. This section is based on the appendix to the paper of Kokic et al. (2011). The linear mixed-effect state-space model generalises the linear

state-space model via the addition of the fixed effect term in the observation equation. This means that the methods used for linear state-space parameter estimation can be adjusted to allow parameter estimation in the LMESS framework. In this paper, I will use an EM algorithm approach, iterating between a Kalman smoother to estimate the unobserved state variable α_{jt} and analytically derived maximum likelihood estimates. First, I explain the Kalman filter and smoother in the E-step detailing the difference with the conventional filter and smoother. Thereafter, I give the explicit formulas derived from the likelihood function used in the M-step. For notation convenience, the subscript j is dropped in this section, as parameter estimation is done separately at each measurement site j.

The parameters that provide the best model fit given the data are found by maximising the likelihood function of the LMESS model. As maximising the likelihood is equivalent to maximising the log-likelihood and the latter is computationally less intensive, the EM algorithm aims to find parameters such that the log-likelihood is maximised. Suppose we already know which covariates are fixed and random effects, so the matrices X_t and Z_t are known for all t. If the initial state is assumed normally distributed with mean π_1 and variance Σ_1, the joint log-likelihood of the observations y and states α is

$$\begin{aligned}\log L(\alpha, y) = &- \sum_{t=1}^{T} \left(\frac{1}{2}(y_t - Z_t\alpha_t - X_t\beta)' R^{-1}(y_t - Z_t\alpha_t - X_t\beta) \right) - \frac{T}{2}\log|R| \\ &- \sum_{t=2}^{T} \left(\frac{1}{2}(\alpha_t - A\alpha_{t-1})' Q^{-1}(\alpha_t - A\alpha_{t-1}) \right) - \frac{T-1}{2}\log|Q| \\ &- \frac{1}{2}(\alpha_1 - \pi_1)'\Sigma_1^{-1}(\alpha_1 - \pi_1) - \frac{1}{2}\log|\Sigma_1| - Tk\log(2\pi).\end{aligned} \tag{4}$$

The EM algorithm works via calculation of state estimates conditional on all data y in the E-step. Then, estimates for A, Q, R, and β are given by analytical solutions to the maximisation problem in the M-step, using the state estimates from the E-step. The E- and M-step are iterated until convergence.

2.2.1 Filtering and smoothing the LMESS model

The E-step finds estimates for the unobserved state variables α_t based on the full set of observations $\mathbf{y}_T$. Only estimating the states however does not suffice, as there are also vector products to be considered in equation (4). For the full E-step we thus need to

estimate three sufficient statistics, given by

$$\hat{\alpha}_{t|T} = \mathbb{E}(\alpha_t|\mathbf{y}_T), \qquad \hat{P}_{t|T} = \mathbb{E}(\alpha_t\alpha_t'|\mathbf{y}_T), \qquad \text{and} \quad \hat{P}_{t,t-1|T} = \mathbb{E}(\alpha_t\alpha_{t-1}'|\mathbf{y}_T),$$

where $\hat{\alpha}_{t|T}$ is the smoothed state vector. Both $\hat{P}_{t|T}$ and $\hat{P}_{t,t-1|T}$ are related to the smoothed state variance and smoothed transition variance, but have no physical interpretation. From these three statistics, I can also calculate the smoothed state variance $\hat{\Sigma}_{t|T} = \hat{P}_{t|T} - \hat{\alpha}_{t|T}\hat{\alpha}_{t|T}'$.

First, the Kalman filter uses observations up to time t to give initial estimates for the filtered states $\hat{\alpha}_{t|t}$ and filtered covariance $\hat{\Sigma}_{t|t}$. The equations to find these estimates are very similar to the equations for the linear state-space model, but there is an extra term corresponding to the fixed effects added in the equations. The Kalman filter equations for the LMESS model are then given by

$$\begin{aligned} \hat{\alpha}_{t|t} &= A\hat{\alpha}_{t-1|t-1} + K_t e_t & V_t &= A\hat{\Sigma}_{t-1|t-1}A' + Q \\ \hat{\Sigma}_{t|t} &= V_t - K_t Z_t V_t & K_t &= V_t Z_t'(R + Z_t V_t Z_t')^{-1} \\ & & e_t &= y_t - Z_t A\hat{\alpha}_{t-1|t-1} - X_t\beta. \end{aligned}$$

The matrix V_t is the forecasted state covariance matrix for state α_t based on the observations up to $t-1$. The matrix K_t is included to improve computational efficiency, but has no interpretation. The error e_t represents the deviation between the true observation y_t and the expected value based on the observations up to $t-1$. The matrices A, Q, and R, as well as the vector β are maximum likelihood estimates from the previous M-step in the algorithm.

When the filter has finished, the Kalman smoother gives estimates for the state based on the full sample period. The Kalman smoother equations for the linear mixed-effect state-space model are equal to the Kalman smoother equations for the linear state-space model. We run the smoother backwards, initialising $\hat{\alpha}_{T|T} = \hat{\alpha}_{T|t=T}$, where $\hat{\alpha}_{t|T}$ is the smoothed state at time t. Furthermore, this is also the step where I estimate the sufficient statistics needed. The full set of smoother equations is

$$\begin{aligned} \hat{\alpha}_{t-1|T} &= \hat{\alpha}_{t-1|t-1} + \hat{\Sigma}_{t-1|t-1}A'V_t^{-1}(\hat{\alpha}_{t|T} - \hat{\alpha}_{t|t}) \\ \hat{\Sigma}_{t-1|T} &= \hat{\Sigma}_{t-1|t-1} - \hat{\Sigma}_{t-1|t-1}A'V_t^{-1}(V_t - \hat{\Sigma}_{t|T})V_t^{-1}A\hat{\Sigma}_{t-1|t-1} \\ \hat{P}_{t|T} &= \hat{\Sigma}_{t|T} + \hat{\alpha}_{t|T}\hat{\alpha}_{t|T}' \\ \hat{P}_{t,t-1|T} &= \hat{\Sigma}_{t|T}V_t^{-1}A\hat{\Sigma}_{t-1|t-1} + \hat{\alpha}_{t|T}\hat{\alpha}_{t-1|T}', \end{aligned}$$

where $\hat{\alpha}_{t|t}$, $\hat{\Sigma}_{t|t}$, and V_t are taken from the Kalman filter and A is the maximum likelihood estimate from the previous M-step in the algorithm.

2.2.2 Maximising the log-likelihood

In this research, I assume normality, which leads to the log-likelihood function in equation (4). Maximisation equations for each of the model parameters are then found by calculating the partial derivatives for the conditional log-likelihood on all observations. For example, the partial derivative with respect to the transition matrix A is given by

$$\frac{\partial\ \mathbb{E}(\log L(\alpha, y)|\mathbf{y}_T)}{\partial A} = -\sum_{t=2}^{T} Q^{-1}\hat{P}_{t,t-1|T} + \sum_{t=2}^{T} Q^{-1}A\hat{P}_{t-1|T}.$$

At the value of maximum likelihood, the expression on the right hand side must equal zero. After rewriting this equation by cancelling the inverted matrix Q and isolating the term A, the maximum likelihood estimate is given by

$$\hat{A} = \sum_{t=2}^{T} \hat{P}_{t,t-1|T}\left(\sum_{t=2}^{T} \hat{P}_{t-1|T}\right)^{-1}. \tag{5}$$

Applying the same approach of analytical derivation to the other model parameters β, R, and Q, gives us the following set of maximum likelihood estimates.

$$\hat{\beta} = \left(\sum_{t=1}^{T} X_t'\hat{R}^{-1}X_t\right)^{-1} \sum_{t=1}^{T} X_t'\hat{R}^{-1}(y_t - Z_t\hat{\alpha}_{t|T}) \tag{6}$$

$$\hat{R} = T^{-1}\sum_{t=1}^{T}\left[(y_t - Z_t\hat{\alpha}_{t|T} - X_t\hat{\beta})(y_t - Z_t\hat{\alpha}_{t|T} - X_t\hat{\beta})' + Z_t\hat{\Sigma}_{t|T}Z_t'\right] \tag{7}$$

$$\hat{Q} = (T-1)^{-1}\sum_{t=2}^{T}\left[\hat{P}_t - \hat{A}\hat{P}_{t,t-1}' - \hat{P}_{t,t-1}\hat{A}' + \hat{A}\hat{P}_{t-1}\hat{A}'\right] \tag{8}$$

The set of estimates in equations (5) to (8) forms the M-step of the parameter estimation procedure for the linear mixed-effect state-space model.

2.3 Variable selection

The second problem posed by the linear mixed-effect state-space model is the selection of covariates to include as fixed or random effects for each dependent variable. In this research, I investigate four dependent variables and six explanatory covariates. As each covariate can be included as fixed or random effect or excluded, there are $3^{6\cdot4}$ possible models in the setup of this thesis. Full estimation of over 280 billion possible models would be highly inefficient, if not infeasible. In this section I introduce three methods for variable selection, increasing in complexity and computational demand. Firstly, the

manual stepwise selection procedure used by Kokic et al. (2011) is described. Secondly, the joint Kalman smoother estimation with fixed parameters as described by Chow (1984) and Durbin & Koopman (2012) is translated to application for variable selection in the linear mixed-effect state-space model. Thirdly, the joint selection algorithm proposed by Bondell et al. (2010) is translated to the context of the LMESS model.

2.3.1 Stepwise selection

Kokic et al. (2011) use a manual selection procedure to select which covariates are part of fixed effects matrix X_{ij}, and which are part of random effects matrix Z_{ij}. A priori I assume no knowledge on the true nature of each covariate, so all are included at first. Each time the LMESS model is fitted to a certain set of covariates, I use the estimation equations as detailed in section 2.2. For any dependent variable i at any measurement site j the following procedure is applied:

1. First a multivariate linear regression model is fitted to the time series via a backwards elimination, or *general-to-specific*, procedure, which is initialised with all covariates as possible explanatory variables. Only covariates with at least 10% significance are retained (Heij, De Boer, Franses, Kloek, & Van Dijk, 2004).
2. All retained covariates are modelled as random effect state-space terms. At each time t, the 90% confidence interval is calculated using the smoothed state variance estimate. If the minimum across the upper bounds is smaller than the maximum across the lower bounds, I consider the coefficient to show significant dynamics. Then the covariate is retained as a random effect. If all upper bounds are larger than all lower bounds, the dynamics of the coefficient are not significant and I regard the covariate as a fixed effect.
3. With the new distinction between fixed and random effects, the LMESS model is refitted again and all fixed effects which are not significant at the 10% level are removed.
4. The last step is to test each of the covariates removed in step 1 one-by-one as random effects. Just as in step 2, if the minimum of the 90% confidence interval upper bounds is smaller than the maximum of the lower bounds, I retain the covariate as random effect. If all upper bounds are larger than all lower bounds, I refit the model to include the covariate as a fixed effect. The covariate is retained in the model as fixed

if the coefficient is significant at the 10% level.

Kokic et al. (2011) motivate their selection algorithm for its ability to include covariates which are fixed in the multivariate linear regression model as random. Furthermore, all covariates excluded by the multivariate linear regression in step 1 can still be included in the LMESS model via step 4. The main advantage of this procedure is that there are no restricting assumptions on model specifications. The same dependent variable at different measurement sites can have different covariates as random and fixed effects. The resulting downside is that the full procedure needs to be repeated $r \cdot m$ times. For larger datasets of monthly or daily data, this can be computationally inefficient, as the LMESS model needs to be fitted repeatedly in the algorithm of Kokic et al. (2011). Besides the computational demand, Kokic et al. (2011) also state that their procedure might lead to suboptimal models. This statement is supported by Bondell et al. (2010), who find stepwise algorithms to be biased by the order of selection. Therefore, I consider two models that identify fixed and random effects simultaneously. I include the selection method of Kokic et al. (2011) to serve as a benchmark algorithm to improve upon.

2.3.2 Chow's adjusted Kalman Smoother

The second selection algorithm is based on a state-space formulation first proposed by Chow (1984). The rationale behind this procedure is to estimate all covariates as both fixed and random at the same time and select which covariates to include in the model. It has not yet been implemented in the context of linear mixed-effect variable selection.

In the model formulation of Chow (1984), the researcher would assume a subset of explanatory variables to have a fixed effect. To lower the number of parameters to estimate, the state vector and the corresponding transition matrix are restricted. Durbin & Koopman (2012) specify the parameter estimation for this model formulation in more detail. To use the same algorithm for variable selection, the linear mixed-effect state-space model in equations (1) and (2) needs to be rewritten in a form coherent with the formulation of Chow (1984). As I assume no a priori knowledge on the fixed and random effects, I set $Z_{ijt} = X_{ijt}$ for all (i, j, t) and include all covariates in each row of both matrices. For each dependent variable $i \in \{1, \ldots, r\}$ at each measurement site $j \in \{1, \ldots, m\}$, the

LMESS model can be written as

$$y_{ijt} = X^*_{ijt} \begin{bmatrix} \beta_{ijt} \\ \alpha_{ijt} \end{bmatrix} + v_{ijt} \tag{9}$$

$$\begin{bmatrix} \beta_{ijt} \\ \alpha_{ijt} \end{bmatrix} = B_{ij} \begin{bmatrix} \beta_{ijt-1} \\ \alpha_{ijt-1} \end{bmatrix} + \begin{bmatrix} 0 \\ w_{ijt} \end{bmatrix}, \tag{10}$$

where $X^*_{ijt} = (X_{ijt}, X_{ijt})$ and B_{ij} is a $(2q_{ij} \times 2q_{ij})$ block matrix containing the identity matrix of size q_{ij} in the top left and a $(q_{ij} \times q_{ij})$ transition matrix in the bottom right. v_{ijt} and w_{ijt} are normally distributed errors with mean zero and covariance R_{ij} and Q_{ij} respectively. Just as Durbin & Koopman (2012), I attach subscript notation to β_{ijt} for convenience in the state-space formulation, but note that $\beta_{ijt} = \beta_{ijt-1} = \beta_{ij}$. In equation (9), the structure of the covariates matrix X^*_{ijt} implies that each covariate is related to both a coefficient in β_{ijt} and a component in α_{ijt}.

Estimation of parameters in the model formulation of Chow can be done by restriction of the standard EM-algorithm approach proposed by Shumway & Stoffer (1982). Both E- and M-step are detailed in appendix A. For a given dependent variable i at measurement site j, I use the following procedure to decide which variables to include as fixed and random effects.

1. The state vector and transition matrices are jointly modelled via the Kalman smoother EM algorithm from Shumway & Stoffer (1982). The algorithm is adjusted for notation, the result of which is detailed in appendix A.
2. As in step 2 of the algorithm of Kokic et al. (2011), the 90% confidence intervals for α_{ijt} are calculated for all times t. If the minimum across the upper bounds is smaller than the maximum of the lower bounds, I consider the coefficient to show significant dynamics. Then the covariate is considered a random effect.
3. If all upper bounds of the 90% confidence intervals are larger than all lower bounds, the dynamics of the coefficient are not significant. Then I test whether the coefficient in β_{ij} corresponding to the covariate is significant at the 10% level. If so, the covariate is included as a fixed effect. A covariate is thus omitted from the model if the state-space coefficient shows no significant dynamics and the corresponding coefficient in β_{ij} is not significant at the 10% level.

This procedure needs to evaluate $r \cdot m$ EM-algorithms to find all models. So the number

of EM-algorithms is lower compared to the method proposed by Kokic et al. (2011), but the dimensionality for each algorithm is higher.

2.3.3 Joint selection of correlated data

The algorithm of Bondell et al. (2010) is based on the conventional linear mixed-effect model form, which has T rows corresponding to the observations of the dependent variable and explanatory covariates. Therefore, the state-space formulation in equation (1) is not suitable for their approach. Additionally, to apply the algorithm of Bondell et al. (2010) to the LMESS model, it is assumed that the division of covariates into fixed and random effects at all m measurement sites is the same and the fixed effect coefficient vector β_i is the same at all sites j. For variable i at site j, the linear mixed-effect model is written as

$$y_{ij} = X_{ij}\beta_i + Z_{ij}\bar{\alpha}^*_{ij} + \varepsilon_{ij}, \tag{11}$$

where $y_{ij} = (y_{ij1}, \ldots, y_{ijT})'$, $X_{ij} = (X'_{ij1}, \ldots, X'_{ijT})'$, and $Z_{ij} = (Z'_{ij1}, \ldots, Z'_{ijT})'$. As I assume no a priori knowledge on which covariates are fixed or random, I include all covariates in both X_{ij} and Z_{ij}, which have p_i and q_i columns respectively. The vector $\bar{\alpha}^*_{ij}$ can be seen as a time average random effect, which is assumed to be normally distributed as $N(0, \sigma_i^2\Psi_i)$. I may assume zero-mean, since covariates can be both fixed and random in the formulation of Bondell et al. (2010). Therefore, any non-zero mean random effect will be reflected by a significant coefficient in β_i. The error ε_{ij} is assumed normal, satisfying $\varepsilon_{ij} \sim N(0, \sigma_i^2\Omega_{ij})$. As Pourahmadi & Daniels (2002) show, the implied relation between consecutive measurements in equation (2) can be incorporated in the covariance structure of Ω_{ij}.

Durbin & Koopman (2012) provide a framework for estimation of the full covariance matrix Ω_{ij}, which I adjust to fit with the model formulation as in equations (9) and (10). Then I find

$$\mathrm{Var}(y_{ij}) = \Omega_{ij} = X^*_{ij}B^*_{ij}Q^*_{ij}B^{*\prime}_{ij}X^{*\prime}_{ij} + R^*_{ij}, \tag{12}$$

in which

$$X^*_{ij} = \begin{bmatrix} X^*_{ij1} & & 0 & 0 \\ & \ddots & & \vdots \\ 0 & & X^*_{ijT} & 0 \end{bmatrix}, \qquad R^*_{ij} = \begin{bmatrix} R_{ij} & & 0 \\ & \ddots & \\ 0 & & R_{ij} \end{bmatrix},$$

$$Q^*_{ij} = \begin{bmatrix} P_{ij1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & Q_{ij} & & 0 & 0 \\ \vdots & & & \ddots & & \\ 0 & 0 & 0 & & 0 & 0 \\ 0 & 0 & 0 & & 0 & Q_{ij} \end{bmatrix}, \text{and } B^*_{ij} = \begin{bmatrix} I_{2q_i} & 0 & \cdots & 0 & 0 \\ B_{ij} & I_{2q_i} & & 0 & 0 \\ B^2_{ij} & B_{ij} & & 0 & 0 \\ \vdots & & \ddots & & \\ B^{T-1}_{ij} & B^{T-2}_{ij} & & I_{2q_i} & 0 \\ B^T_{ij} & B^{T-1}_{ij} & \cdots & B_{ij} & I_{2q_i} \end{bmatrix},$$

where P_{ij1} is the smoothed covariance estimate for the initial state vector. As this estimation of the variance is based on the formulation of Chow (1984), I use the same EM algorithm of Shumway & Stoffer (1982) to estimate the parameters in the model.

In order to select which covariates to incorporate as fixed or random, Bondell et al. (2010) factorise the matrix Ψ_i via a modified Cholesky decomposition first described by Chen & Dunson (2003), namely $\Psi_i = D_i\Gamma_i\Gamma_i'D_i$, where D_i is a diagonal matrix and Γ_i a lower triangular matrix with ones on the diagonal. The model in equation (11) may be rewritten as

$$y_{ij} = X_{ij}\beta_i + Z_{ij}D_i\Gamma_i\bar{\alpha}_{ij} + \varepsilon_{ij}, \tag{13}$$

where $\bar{\alpha}_{ij}$ satisfies $\bar{\alpha}_{ij} \sim N(0, \sigma_i^2 I_{q_i})$. Now define the vectors $d_i = (d_{i1}, \ldots, d_{iq_i})'$ and $\gamma_i = (\gamma_{i,kl} : k = 1, \ldots, q_i : l = k+1, \ldots, q_i)'$ containing the free elements of D_i and Γ_i, respectively. With this decomposition, if $d_{in} = 0$ for any n, this is equivalent to removing the n^{th} row and column of the matrix Ψ_i, thus excluding covariate u_n as a random effect for the time series of dependent variable i. Now define the variable $\phi_i = (\beta_i', d_i', \gamma_i')'$ containing all information on the inclusion of fixed and random effects for variable i.

Conditional on X_{ij} and Z_{ij}, the distribution of y_{ij} is normal with mean $X_{ij}\beta_i$ and variance $V_{ij} = \sigma_i^2(Z_{ij}D_i\Gamma_i\Gamma_i'D_iZ_{ij}' + \Omega_{ij})$. After dropping constants, the log-likelihood function as a function of the parameter ϕ_i is written as

$$\mathcal{L}(\phi_i) = -\frac{1}{2}\log|\tilde{V}_i| - \frac{1}{2}(y_i - X_i\beta_i)'\tilde{V}_i^{-1}(y_i - X_i\beta_i), \tag{14}$$

where $\tilde{V}_i = \text{Diag}(V_{i1}, \ldots, V_{im})$, a block diagonal matrix of V_{ij}, and $y_i = (y_{i1}', \ldots, y_{im}')'$ and $X_i = (X_{i1}', \ldots, X_{im}')'$ are the stacked y_{ij} and X_{ij} respectively. The optimal set of variables

ϕ_i is found by maximising the conditional expectation of this log-likelihood along with a penalty function on β_i and d_i. Bondell et al. (2010) proposed an EM-algorithm for this procedure, the expectation step of which is given by

$$g(\phi_i|\phi_i^{(\omega)}) = \mathbb{E}_{\bar{\alpha}_i|y_i,\phi_i}\Big\{||\tilde{\Omega}_i^{-1/2}y_i - \tilde{\Omega}_i^{-1/2}X_i\beta_i - \tilde{\Omega}_i^{-1/2}Z_i\text{diag}(\tilde{\Gamma}_i\bar{\alpha}_i)(\mathbf{1}_m \otimes I_{q_i})d_i||^2\Big\} + \lambda_m\Bigg(\sum_{n=1}^{p_i}\frac{|\beta_{in}|}{|\bar{\beta}_{in}|} + \sum_{n=1}^{q_i}\frac{|d_{in}|}{|\bar{d}_{in}|}\Bigg), \tag{15}$$

in which $\tilde{\Gamma}_i = I_m \otimes \Gamma_i$, $\mathbf{1}_m$ a $(m \times 1)$ vector of ones, and λ_m represents the non-negative regularisation parameter. The vector $\bar{\alpha}_i = (\bar{\alpha}_{i1}', \ldots, \bar{\alpha}_{im}')'$ is the stacked version of all random effect parameters for all different sites. Matrix $\tilde{\Omega}_i = \text{Diag}(\Omega_{i1}, \ldots, \Omega_{im})$ is the block diagonal matrix of covariance matrices. The vector $\bar{\beta}_i$ is the GLS estimate for β_i and $\bar{d}_i$ is found by decomposition of the restricted maximum likelihood variance estimate for Ψ_i. For the M-step, the expression in Equation (15) is minimised over ϕ_i. By iterating between the quadratic programming problem for the vector $(\beta_i', d_i')'$ and the closed form solution for γ_i found by Bondell et al. (2010), we find an updated vector $\phi_i^{(\omega+1)}$ upon convergence. The algorithm is given in more detail in Appendix B. When $\phi_i^{(\omega)}$ has converged, the solution gives the final parameter estimates.

This optimisation is performed for different values of the penalty parameter λ_m. The set of fixed and random effects included in the estimation model is given by the solution that minimises the BIC_{λ_m} criterion given by

$$BIC_{\lambda_m} = -2\mathcal{L}(\hat{\phi}_i) + \log(mT) \cdot df_{\lambda_m}, \tag{16}$$

where $\mathcal{L}(\hat{\phi}_i)$ is the log-likelihood function defined in equation (14) and df_{λ_m} is the number of non-zero elements of the vector $\hat{\phi}_i$. All q_i^* covariates corresponding to non-zero elements in d_i are included as random effects in Z_{jt} in equation (3), and any covariate not included in Z_{jt} with a non-zero coefficient in β_i is regarded as a fixed effect.

A drawback of the method of Bondell et al. (2010) is that it assumes the same model is true across all measurement sites, which may not be valid in all cases. Furthermore, the estimation of the matrix Ω_{ij} relies on state-space estimates. The extra step taken to find the covariance matrices adds another source of uncertainty in the model, which may lead to wrong identification of random and fixed parameters.

2.4 Forecasting

After obtaining the expectation maximisation estimates for the linear mixed-effect state-space model parameters at any site j, I simulate forecasts as follows. For $t > T$, the conditional distributions for the state variables and climate variables are $(\alpha_{jt}|\mathbf{y}_{jT}) \sim N(\hat{\alpha}_{jt}, \hat{\Sigma}_{jt})$ and $(y_{jt}|\mathbf{y}_{jT}) \sim N(\hat{y}_t, Z'_{jt}\hat{\Sigma}_{jt}Z_{jt} + R_j)$, respectively, in which

$$\hat{\alpha}_{jt} = A_j\hat{\alpha}_{jt-1} \tag{17}$$

$$\hat{\Sigma}_{jt} = A_j\hat{\Sigma}_{jt-1}A'_j + Q_j \tag{18}$$

$$\hat{y}_{jt} = X_{jt}\beta_j + Z_{jt}\hat{\alpha}_{jt}. \tag{19}$$

Since the LMESS is a generalisation of a multivariate linear regression model, the added random effects should improve forecast accuracy. To compare the methods, I fit a multivariate linear regression model to the data series and only select covariate terms with at least 10% statistical significance with the backward selection procedure described by Heij et al. (2004). Both LMESS and MLR are compared to a naive long-term mean forecast in terms of the root mean squared forecast error. When forecasting, we assume perfect foresight on the covariates and therefore use the realised values since forecasting of the covariates is beyond the scope of this research.

3 Simulation Study

In this section, I examine the accuracy of the three variable selection algorithms in a simulation study, wherein the ability to find the full correct model is evaluated in a variety of scenarios. Thereafter, I study the influence of wrong model selection on predictive accuracy.

3.1 Variable selection accuracy

The first subject I investigate is whether the three different proposed selection methods in section 2.3 are able to select the correct model specification from simulated data. In this section, I produce artificial data under distinct model assumptions to investigate how different data characteristics influence the accuracy of the three proposed algorithms. The data generating process is a simple linear mixed-effect state-space model, from which the observation equation is given as

$$\begin{bmatrix} y_{1jt} \\ y_{2jt} \end{bmatrix} = \begin{bmatrix} u_{1jt} & u_{2jt} & 0 & 0 \\ 0 & 0 & u_{1jt} & u_{3jt} \end{bmatrix} \beta + \begin{bmatrix} 1 & u_{3jt} & 0 & 0 \\ 0 & 0 & 1 & u_{2jt} \end{bmatrix} \alpha_{jt} + v_{jt}. \tag{20}$$

The number of covariates in this case is thus four including the intercept. The reason to choose this model formulation is that it allows for investigation of data characteristics while still having a low number of parameters to estimate. For both dependent variables y_{1jt} and y_{2jt}, I choose a common random effect in the form of the intercept and a common fixed effect in the form of u_{1jt}. The covariate u_{2jt} is included as a fixed effect for y_{1jt} and as a random effect for y_{2jt}, whereas this is reversed for covariate u_{3jt}. The last two covariates are included to investigate the influence of data characteristics on model selection accuracy. Suppose the covariate u_{3jt} has a dominant seasonality component which limits the accuracy of all three selection algorithms if it is included as random effect but not if it is included as fixed. This would lead to a distinct difference between y_{1jt} and y_{2jt} in terms of model selection accuracy. In theory, in my simulation study, this would lead to a lower accuracy in y_{1jt} compared to y_{2jt}.

I investigate the influence of all model parameters on forecast accuracy as follows. First, a baseline scenario is considered and all three proposed variable selection algorithms are tested on their accuracy. After the initial analysis, I construct new data by changing one model parameter ceteris paribus and apply all three selection algorithms. The baseline

scenario considers a situation with $m = 5$ measurement sites, each having $T = 100$ observations y_{jt}. The covariate u_{1jt} is generated from a uniform $(-1, 1)$ distribution. To investigate the influence of a trending covariate by generating u_{2jt} from the addition of a linear trend $0.05t - 2.5$ and a uniform $(-0.5, 0.5)$ distribution. And to research the influence of a seasonal covariate, I simulate u_{3jt} from the addition of a cosine function $1.25\cos(\pi \cdot t/6)$ and a uniform $(-0.75, 0.75)$ distribution. The fixed effect coefficient in the baseline scenario is a vector of ones, so $\beta = (1, 1, 1, 1)'$. The initial state vector α_{j1} determines how large the random effect coefficients are and in the baseline case is chosen to give values of similar size to β. In this first scenario, I draw the initial state from

$$\alpha_{j1} \sim N\left(\begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \ \mathrm{diag}(\sigma_{\alpha_1}) \cdot \begin{bmatrix}1 & 0.45 & 0 & 0\\ 0.45 & 1 & 0 & 0\\ 0 & 0 & 1 & 0.45\\ 0 & 0 & 0.45 & 1\end{bmatrix} \cdot \mathrm{diag}(\sigma_{\alpha_1})\right),$$

with $\sigma_{\alpha_1} = (1, 1, 1, 1)'$ the vector of standard deviations. The state transition matrix A_j is a diagonal matrix with non-zero elements drawn from a uniform $(0.75, 1)$ distribution. The observation equation error v_{jt} is normally distributed with mean zero and variance $R_j = 0.4 \cdot I_2$. Similarly, the state equation error w_{jt} is normally distributed with mean zero and variance $Q_j = 0.3 \cdot I_4$.

To evaluate the model selection accuracy of each algorithm, I consider the following other scenarios. I take a different number of observations per measurement site with $T = 50$ and $T = 200$. I increase in the number of measurement sites to $m = 10$ and $m = 20$. I also change the ratio between the fixed and random effect coefficients by setting $\sigma_{\alpha_1} = (3, 2, 3, 2)'$ or $\beta = (5, 3, 2, 4)'$. I also vary the signal-to-noise ratio in both the observation and state equation by considering $R_j = I_2$ and $Q_j = I_4$. Lastly, I evaluate how well the algorithms perform if all covariates are randomly drawn from either a uniform $(-2, 2)$ or a standard normal distribution. In all scenarios, I generate 100 datasets, which means there are $100m$ models for the algorithm of Kokic et al. (2011) and my proposed method based on Chow (1984) to evaluate, as these algorithms allow for the fixed and random effects to be different for different measurement sites. The algorithm by Bondell et al. (2010) has 100 models to find the fixed and random effects for regardless of the number of measurement sites, as it assumes all sites to have the same fixed and random effects a priori.

Table 1 presents the percentage of times the correct model is identified in 100 simulated datasets for each of the three selection algorithms across eleven scenarios. Columns one and four show that the algorithm of Kokic et al. (2011) is suboptimal in its selection accuracy, resulting in correct identification percentages below 20% in all eleven scenarios across both variables. The lowest percentage of correct identification for variable y_1 for the selection algorithm of Kokic et al. (2011) is observed in the second scenario, where the number of observations is smallest. The small amount of observations may lead to over-fitting of the LMESS model parameters in different stages of the selection algorithm. This assumption is supported by the finding that the highest accuracy of 20% is reported for the scenario where $T = 200$. As the number of observations per measurement site increases, the algorithm of Kokic et al. (2011) performs better, but never exceeds 20% accuracy. However, the findings for y_2 don't support this finding as both cases show equal accuracy of 7.2%. Apart from the scenarios with a different number of observations, the algorithm of Kokic et al. (2011) is consistent across different model parameters, both for y_1 and y_2. However, it is never the most accurate across all three selection algorithms. The limited accuracy of their method is in line with the remarks of Kokic et al. (2011), who stated that their method would likely be suboptimal. The findings of Bondell et al. (2010) support this conclusion, as they also find step-wise selection methods to be suboptimal.

In all eleven scenarios, the algorithm based on the formulation of Chow (1984) is most consistent across the variables y_1 and y_2. Table 1 shows that my proposed method outperforms the other two algorithms in 20 out of 22 cases when looking at full model selection. The accuracy for both variables is lowest in the second scenario, where the accuracy drops to 31.6% and 25.8% for y_1 and y_2 respectively. This reduction is due to the limited number of observations T, which leads to overfitting of model parameters, resulting in wrong estimates. The other significant drop in model selection accuracy is in the case where the observation equation covariance is increased. In this scenario, the signal-to-noise ratio in the observation equation is lower than the baseline scenario, which makes it harder for the algorithm to identify the underlying process. The accuracy of the algorithm based on Chow (1984) significantly increases when the seasonality and trending behaviour is removed from the covariates. In the last two scenarios, when all covariates are drawn from standard distribution, the selection accuracy increases to around 70%. Besides the four exceptions, the algorithm by Chow (1984) is consistent across all scenarios,

Table 1

Selecting The Full Model

This table gives the number of times (in %) the covariates were correctly identified with each of the three variable selection methods in 100 simulated datasets, with a data generating process as in equation (20). Column one gives the variable that is changed compared to the baseline* scenario. Columns two to four give the percentage of cases where the model was correctly identified for variable y_1. Columns five to seven give the percentage of cases where the model was correctly identified for variable y_2. Across all three model selection algorithms, the most accurate is marked in bold font.

	y_1			y_2		
Scenario	Kokic	Chow	Bondell	Kokic	Chow	Bondell
Baseline*	13.6	**56.2**	6.0	11.0	**54.6**	50.0
$T = 50$	3.2	**31.6**	12.0	7.2	25.8	**32.0**
$T = 200$	20.0	**59.4**	0.0	7.2	**59.8**	0.0
$m = 10$	13.7	**54.4**	0.0	9.4	**52.9**	0.0
$m = 20$	14.5	**56.4**	0.0	8.4	**54.6**	0.0
$\sigma_{\alpha_1} = (3, 2, 3, 2)'$	12.4	**58.4**	3.0	5.4	52.2	**76.0**
$\beta = (5, 3, 2, 4)'$	10.0	**60.0**	0.0	9.0	**56.4**	23.0
$R_j = I_2$	16.2	**44.8**	2.0	9.2	**39.8**	26.0
$Q_j = I_4$	10.2	**53.6**	6.0	9.4	**63.6**	41.0
$u_{kt} \sim U(-2, 2)$	10.8	**68.0**	9.0	11.6	**70.6**	19.0
$u_{kt} \sim N(0, 1)$	9.8	**70.4**	16.0	7.8	**72.8**	16.0

* The baseline scenario has the following model parameters: $T = 100$, $m = 5$, $\sigma_\alpha = (1, 1, 1, 1)'$, $\beta = (1, 1, 1, 1)'$, $R_j = 0.4I_2$, $Q_j = 0.3I_2$. The covariates are generated from $u_{1t} \sim U(-1, 1)$, $u_{2t} \sim (U(-0.5, 0.5) + 0.05 \cdot (t - 0.5T))$, and $u_{3t} \sim (U(-0.75, 0.75) + 1.25\cos(\pi t/6))$.

with accuracy between 50 and 60 percent. Furthermore, the simultaneous method based on Chow (1984) consistently outperforms the stepwise procedure of Kokic et al. (2011) in all eleven simulated scenarios. The algorithm of Bondell et al. (2010) proves to be most accurate of all three model selection algorithms for variable y_2 when $T = 50$ and $\sigma_{\alpha_1} = (3, 2, 3, 2)'$. However, in all other cases it is outperformed by the algorithm based on the formulation of Chow (1984). Out of the three selection methods, the Bondell et al. (2010) algorithm is the most inconsistent. In some cases, the correct model is never identified, whereas the accuracy for variable y_2 in the scenario with $\sigma_{\alpha_1} = (3, 2, 3, 2)'$ is 76%. The accuracy only exceeds 40% in three cases for y_2. For variable y_1, on the other hand, the largest accuracy observed is 16%, and only above ten percent in two out of eleven scenarios.

In conclusion, with this simulation study I have confirmed the remarks of Kokic et al.

(2011) that their variable selection procedure is suboptimal. Two proposed alternatives show better accuracy in at least some cases, of which the algorithm based on the state-space formulation of Chow (1984) is the most reliable across different model characteristics. The only condition is that the number of observations T must be large enough to prevent over-fitting of model parameters. Although the algorithm by Bondell et al. (2010) shows the highest observed accuracy, it is not robust under different simulation setups. In the application of my methods to real data, I will use the algorithm based on Chow (1984) to identify fixed and random effects.

3.2 Influence on forecasting

In this section, I perform a second simulation study to demonstrate the need for an accurate variable selection algorithm. I compare the forecast accuracy under different variable selection outcomes based on three different measures, namely root mean squared forecast error, mean absolute forecast error and a test proposed by Heij et al. (2004). The data is generated via the same process as in the baseline scenario in section 3.1, apart from that I only use $m = 1$ measurement site and change the total number of observations to 200. So, the DGP in this simulation study is a linear mixed-effect state-space model with the observation equation

$$\begin{bmatrix} y_{1t} \\ y_{2t} \end{bmatrix} = \begin{bmatrix} u_{1t} & u_{2t} & 0 & 0 \\ 0 & 0 & u_{1t} & u_{3t} \end{bmatrix} \beta + \begin{bmatrix} 1 & u_{3t} & 0 & 0 \\ 0 & 0 & 1 & u_{2t} \end{bmatrix} \alpha_t + v_t. \tag{21}$$

In this section, the covariate u_{1t} is generated from a uniform $(-1, 1)$ distribution. The covariate u_{2t} serves as a linear trend variable and is obtained from the addition of a linear trend $0.05t - 5$ and random draws from a uniform $(-0.5, 0.5)$ distribution. Lastly, the covariate u_{3t} is used as seasonal variable and is generated from the addition of cosine function $1.25 \cos(\pi \cdot t/6)$ and draws from a uniform $(-0.75, 0.75)$ distribution. The fixed effects parameter β is set as a vector of ones and the initial state vector α_{j1} is drawn from the distribution

$$\alpha_{j1} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.45 & 0 & 0 \\ 0.45 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.45 \\ 0 & 0 & 0.45 & 1 \end{bmatrix} \right).$$

The covariance of the observation and state equation of the simulated data are given by $R = 0.4 \cdot I_2$ and $Q = 0.3 \cdot I_4$, respectively. Within the full time series of length 200, the estimation period has length $T = 150$ and I create two forecast periods of length $T_f = 10$ and $T_f = 50$. I assume perfect foresight on the covariates when forecasting, so I use the realised values for u_{1t}, u_{2t}, and u_{3t} for $t > 150$.

In total, six different model specifications are evaluated on their forecast accuracy, given in column 3 and 4 of table 2. All models are observed outcomes from the algorithms in the simulation study in section 3.1, but do not span all observed outcomes. The first specification represents the true model, where covariates are correctly specified as random and fixed. The second specification is inverted, where the covariates that are random in the DGP are regarded as fixed during estimation and vice versa. Third is a 'wrong' specification, where one random and one fixed effect are switched for each dependent variable. I also examine a case where one covariate is missing from the fixed terms for y_1 and one covariate is missing from the random effects for y_2. Lastly, I examine cases where either all covariates are regarded as fixed terms, which coincides with the multivariate linear regression model, or all covariates are regarded as random, which resembles a standard linear state-space model.

I measure forecast accuracy in three different ways, the root mean squared forecast error (RMSFE), the mean absolute forecast error (MAFE) and a comparative test proposed by Heij et al. (2004), among others. The first two methods are common forecast evaluation methods, which for any variable y_i are given by

$$\text{RMSFE}_i = \left(\frac{1}{T_f} \sum_{h=1}^{T_f} (y_{i,T+h} - \hat{y}_{i,T+h})^2 \right)^{1/2}, \text{ and } \qquad \text{MAFE}_i = \frac{1}{T_f} \sum_{h=1}^{T_f} |y_{i,T+h} - \hat{y}_{i,T+h}|,$$

where $\hat{y}_{i,T+h}$ is the h-step ahead forecast and $y_{i,T+h}$ is the actual observation at time $T+h$. As Heij et al. (2004) propose, two different models can also be compared by the number of times N_f that the absolute error $|y_{i,T+h} - \hat{y}_{i,T+h}|$ of one model is smaller than the absolute error of another model. If the forecasts are equally good, then N_f should follow a binomial distribution with T_f repetitions and probability $\frac{1}{2}$. If I find N_f to have a probability significantly larger than $\frac{1}{2}$, then the first model is preferred, while a value smaller than $\frac{1}{2}$ would indicate that the second model provides better forecasts. As Wackerly, Mendenhall, & Scheaffer (2014) show, if $T_f > 9$, I can approximate the binomial distribution with a normal distribution with mean $\frac{1}{2}T_f$ and variance $\frac{1}{4}T_f$. From the approximate normal

distribution I can derive a test statistic

$$B = \frac{2 \cdot N_f - T_f}{\sqrt{T_f}} \sim N(0,1). \tag{22}$$

The reasoning behind the construction of B is given in appendix C. In my simulation study, the true model with the correct covariates is the baseline to which all other methods are compared and I test the null hypothesis at the 95% confidence level, so I do not reject the null hypothesis if $B \in [-2, 2]$.

Table 2 shows the accuracy of all six covariate selection scenarios in terms of RMSFE, MAFE and, for all but the true model, the performance relative to the true model specification. Rows one to six report the results in forecasting variable y_1, where the true model specification yields the smallest RMSFE and MAFE at the short forecast horizon $T_f = 10$. For a longer forecast period, the true model is not distinctly different from the

Table 2

Forecasting in wrong models

This table reports on the influence of different variable selection outcomes on forecasting accuracy. A 100 datasets are generated via the LMESS model in equation (21), with an estimation period of $T = 150$ and forecasting period of $T_f = 10$ and $T_f = 50$. Columns 3 and 4 report the covariates associated with the dependent variable in the estimation algorithm, where c denotes a constant. Columns 5 and 6 report the average Root Mean Squared Forecast Error (RMSFE) in each specification and columns 7 and 8 report the average Mean Absolute Forecast Error (MAFE) for different forecast horizons. Column 9 reports the percentage of datasets wherein the forecasts by the model are significantly better than the forecasts of the true model and column 10 shows the percentage of datasets wherein the true model forecasts are significantly better.

		Covariates		RMSFE		MAFE		Test (%)	
		Fixed	Random	$T_f = 10$	$T_f = 50$	$T_f = 10$	$T_f = 50$	$B < -2$	$B > 2$
y_1	True	u_1, u_2	c, u_3	**1.87**	3.20	**1.58**	2.70		
	Inverted	c, u_3	u_1, u_2	2.47	3.74	2.12	3.21	18	47
	Wrong	c, u_1	u_2, u_3	2.02	3.19	1.75	2.69	31	41
	Missing	u_1	c, u_3	2.07	3.73	1.80	3.19	21	64
	All fixed	c, u_1, u_2, u_3		1.96	**2.65**	1.66	**2.19**	36	28
	All random		c, u_1, u_2, u_3	2.89	4.08	2.58	3.59	14	69
y_2	True	u_1, u_3	c, u_2	**3.05**	**4.86**	**2.61**	**3.97**		
	Inverted	c, u_2	u_1, u_3	3.96	5.48	3.49	4.56	10	37
	Wrong	u_1, u_2	c, u_3	4.64	8.71	4.07	7.43	7	52
	Missing	u_1, u_3	c	3.09	5.08	2.64	4.16	18	24
	All fixed	c, u_1, u_2, u_3		3.81	5.37	3.37	4.48	17	40
	All random		c, u_1, u_2, u_3	3.83	5.28	3.36	4.37	9	23

'Wrong' specification, with both RMSFE and MAFE values only 0.01 apart. In terms of test statistics, the 'Wrong' model is preferred over the true specification in 31 out of 100 datasets, compared to 41 cases where the true model provides significantly better forecasts. The model with all covariates as fixed effects is the only model with a significantly lower RMSFE and MAFE than the true model. The statistic test results in columns 9 and 10 support these findings, as the model with all covariates as fixed is preferred in 36 out of 100 datasets, whereas the true model is preferred in 28 of the datasets. The 'Inverted', 'Missing', and 'All random' model specifications are all outperformed by the true model, mainly in terms of the statistical test, where the true model is preferred at least twice as often in the 100 simulated datasets.

The results for variable y_2 are reported in rows 7 to 12, and show that the true model is preferred for forecasting relative to all other specifications. Both in terms of RMSFE and MAFE, the true model shows the lowest values of all specifications, with the 'Missing' specification similar for the short forecast horizon. In terms of statistically significant preference, the 'Missing' model is also closest to the true model, being preferred in 18 out of 100 datasets compared to 24 times for the true model. For the other four wrong model forecasts, the results are more pronounced, with the true model being preferred at least twice as often in 100 simulated datasets. The 'Wrong' specification, which shows similar RMSFE and MAFE for variable y_1, shows the worst forecasts for variable y_2 and is least often preferred over the true model in terms of accuracy, only in 7 out of 100 cases compared to 52 cases where the true model is significantly better than the 'Wrong' model.

Out of the five wrong model specifications, none consistently outperforms the true model specification in forecast accuracy across both dependent variables in this simulation study. This shows that selection of the true model influences forecast accuracy significantly. Given the results in section 3.1, the implication of this finding is that forecasting of real time series with the methods described is uncertain. The aggregate inaccuracies due to the chance of wrong model specification and the influence of wrong model specification on forecasting performance, might limit the applicability of my proposed method. These findings are further explored in the next two sections where I apply my proposed variable selection method on climate data in the Netherlands.

4 Data

In this study I use historical monthly data of four dependent variables and five covariates for five weather stations in the Netherlands. The sample period spans from January 1959 to December 2017, thus consisting of 708 monthly observations. The five weather stations from where the data is retrieved are De Bilt, De Kooy, Eelde, Vlissingen and Maastricht/Beek. Figure 1 shows the locations of all five stations within the Netherlands. These five weather stations are used for two reasons. Firstly, they span most regions within the country of the Netherlands, thus providing basis to test whether the assumption of equal model between all measurements sites is valid. Secondly, the time series of these five stations are homogeneously corrected by the KNMI for changing measurement techniques. This correction makes them usable for time series analysis.

The dependent variables of interest are the proportion of days per month with at least 1 mm of rainfall p_{rain}, mean rainfall on wet days μ_{rain}, mean maximum daily temperature $Temp_{max}$, and mean minimum daily temperature $Temp_{min}$. Time series for these four variables were retrieved from Koninklijk Nederlands Meteorologisch Instituut (KNMI) (2018). As I assume normality for the dependent variables, I transform the time series to better fit the model specification. More precisely, I use the logit or log-odds function from the proportion of rainy days and the logarithm of the mean rainfall. For both temperature time series the 12-month difference of monthly averages is calculated to

Figure 1. Weather stations. This figure shows the locations of the five observatories in the Netherlands used in this study.

Table 3

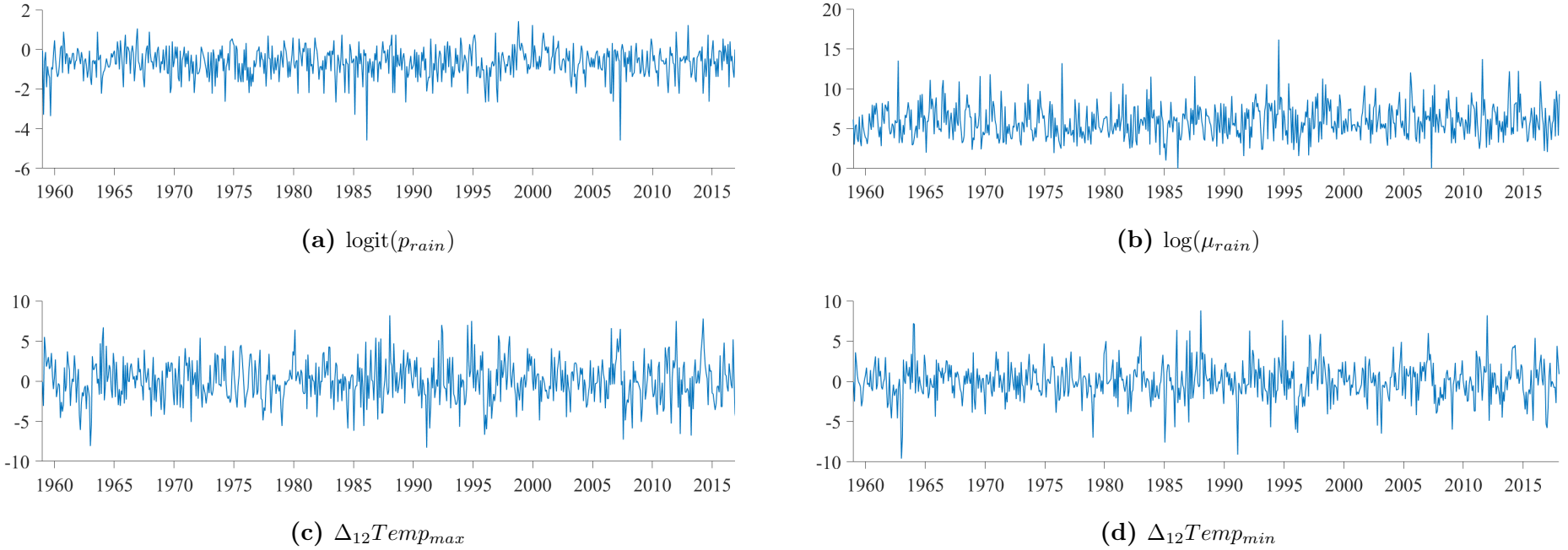
Summary Statistics De Bilt

This table shows the summary statistics for the dependent and explanatory variables used in this study. The values are reported for the weather station in De Bilt in the sample period January 1959 through December 2017. Rows one to four show the pre-transformed dependent variables and rows five to nine show the five explanatory variables considered in this study.

	Mean	Std. Dev.	Min	Max
Dependent variables				
$\text{logit}(p_{rain})$	−0.654	0.753	−4.595	1.427
$\log(\mu_{rain})$	5.916	2.105	0.000	16.167
$\Delta_{12}Temp_{max}$	0.033	2.569	−8.300	8.200
$\Delta_{12}Temp_{min}$	0.015	2.267	−9.600	8.800
Explanatory covariates				
CO_2 level	353.481	26.713	313.260	409.650
NAO index	0.013	1.020	−3.180	3.040
ρ_{cyc}	3.285	1.214	0.000	7.000
$\rho_{anti\text{-}cyc}$	3.480	1.186	1.000	7.000
RH	0.818	0.064	0.620	0.931

remove the dominant seasonality in these time series. Table 3 reports the summary statistics of all four transformed time series for the weather station at De Bilt. Figure 2 shows the time series of the four pre-transformed dependent variables in De Bilt over the full sample period from January 1959 to December 2017.

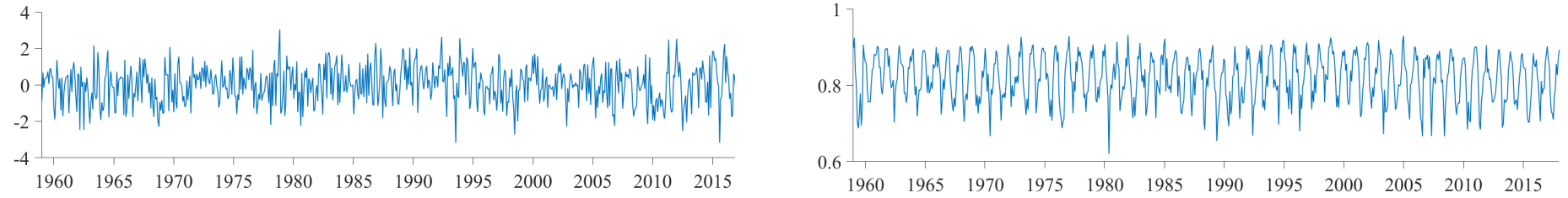
The drivers of weather in Western-Europe have been subject of much research. In this research I investigate the influence of five factors. Two are large scale variables, namely the North Atlantic Oscillation index and the atmospheric level of CO_2. For each measurement site, I also use local indicators, which are the cyclone density, anti-cyclone density, and relative humidity. As the linear mixed-effect state-space model allows for time-varying coefficients, the covariates need not be transformed. Figures 3, 4, and 5 show all time series for the explanatory covariates.



(a) $\text{logit}(p_{rain})$ **(b)** $\log(\mu_{rain})$

(c) $\Delta_{12} Temp_{max}$ **(d)** $\Delta_{12} Temp_{min}$

Figure 2. Dependent variables. This figure shows the pre-transformed dependent variables considered in this study at the weather station at De Bilt over the sample period from January 1959 to December 2017.



(a) NAO Index **(b)** Relative Humidity (RH)

Figure 3. Covariates. This figure shows two explanatory covariates considered in this study over the sample period from January 1959 to December 2017, which are (a) monthly North Atlantic Oscillation (NOA) as defined by the National Weather Service Climate Prediction Center and (b) Relative Humidity at De Bilt.

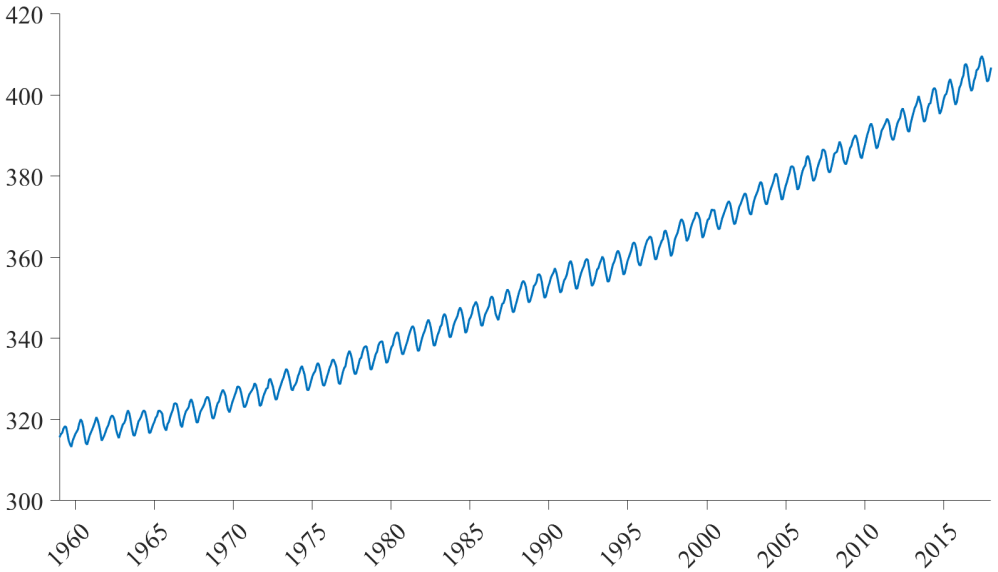
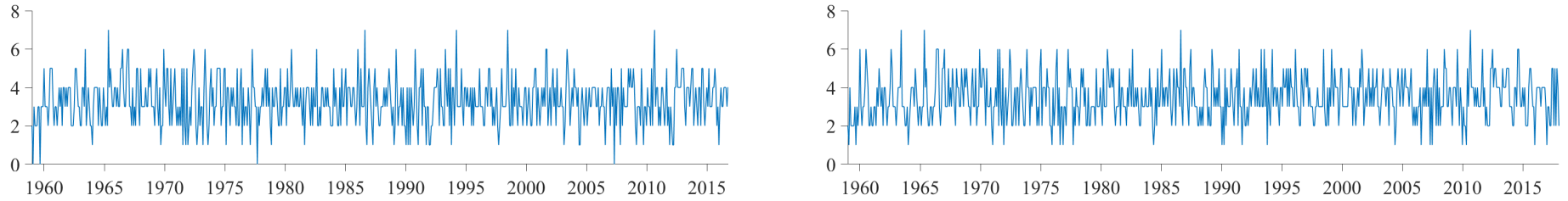


Figure 4. Keeling Curve. This figure shows the monthly average atmospheric level of CO_2 in parts per million (ppm) at the Mauna Loa Observatory in Hawaii during the sample period from January 1959 to December 2017.



(a) Cyclone density (ρ_{cyc})

(b) Anti-cyclone density ($\rho_{anti\text{-}cyc}$)

Figure 5. Covariates. This figure shows two explanatory covariates considered in this study over the sample period from January 1959 to December 2017 at De Bilt, which are (a) the cyclone density and (b) the anti-cyclone density.

Kerr (1997), Hoerling et al. (2001), and Donat, Leckebusch, Pinto, & Ulbrich (2009), among others, show the influence of the North Atlantic Oscillation (NAO) as a main factor on precipitation, temperature and wind in the region. The NAO index depicted in figure 3a is a measure based on the difference of normalised sea level pressure between Lisbon, Portugal and Stykkisholmur/Reykjavik, Iceland (Hurrell, 1995). Although slightly different definitions exist, I use the time-series defined by the National Weather Service Climate Prediction Center in this study, as the difference between definitions is always within 1%.

The atmospheric level of carbon dioxide is another factor widely associated with weather systems across the world. Arrhenius (1896) demonstrates the physics underlying the relation between rising CO_2 levels and increasing global temperatures. The observed temperature development can not be explained without accounting for greenhouse emissions (Pachauri et al., 2014). Figure 4 shows the time series from the Mauna Loa Observatory in Hawaii, which is the longest continuous record of direct atmospheric CO_2 measurements. This time series representation is known as the Keeling Curve, named after Charles David Keeling, who started the measurements in 1958. Not only does figure 4 show the upward trend in atmospheric CO_2 levels, it also shows seasonal patterns, resulting in a sawtooth pattern. Although geographic variations in atmospheric carbon dioxide levels are common, the difference between Mauna Loa and the global average is always within 0.5%, making it a representative time series.

Atmospheric pressure is generally used as indicator of the state of the weather (Neiburger, Edinger, & Bonner, 1971). Low pressure can serve as a proxy for bad weather with lower temperatures and more rainfall, whereas high pressure is associated with higher temperatures and less rainfall. For each month, I calculate the cyclone density ρ_{cyc} by counting the number of periods of at least one day during which the daily atmospheric pressure is below 1013.5 hPa. The value of 1013.5 is chosen as it is the time series average of the data used in the study. Similarly, the anti-cyclone density $\rho_{anti\text{-}cyc}$ is calculated as the number of periods of at least one day during which the daily atmospheric pressure stays above 1013.5 hPa.

The relative humidity RH has also been identified as an important predictor of precipitation in the Netherlands by Beckmann & Buishand (2002). Relative humidity is defined as the partial pressure of water vapour in air divided by the vapour pressure of

water at the ambient temperature (Perry, 1950). I take the monthly average from daily data at each weather station and include it as a predictor. Figure 3b shows the seasonal behaviour of the relative humidity. Each years peak occurs in the winter months between January and March, as the cold air has a lower vapour pressure, thus the air is more saturated. In summertime, when the ambient temperature is higher, the relative humidity is lower due to warmer airs higher capacity to hold water vapour.

5 Application in the Netherlands

In this section, I apply the linear mixed-effect state-space model approach to the data described in section 4. In ten different estimation periods, I compare the forecast accuracy of the LMESS model relative to a climatology forecast for each month. To place the results into context, I calculate the same statistics for a backward selection multivariate linear regression model. This MLR model is obtained by the procedure in step 1 in the selection algorithm of Kokic et al. (2011), described in section 2.3.1.

The first step is to select the covariates to include as random and fixed effects for each dependent variable in the LMESS model at each measurement site. Given the results of the simulation study in section 3, I use the adjusted Chow algorithm for the selection of variables. The estimation period is given by an expanding window starting at January 1959 and ending in December of 1998 for the first case. The estimation period is then extended by 12 months for each new variable selection procedure, such that the last estimation period spans from January 1959 to December 2007. This procedure gives me ten different datasets to evaluate forecast performance. After identifying the random and fixed effects for each dependent variable, I estimate the model parameters via the EM-algorithm of Kokic et al. (2011) described in section 2.2. The next step is to produce monthly forecasts over a 10-year period as described in section 2.4. I use the realised covariate data in the forecast window as covariate estimation is outside of the scope of this research.

A second type of LMESS model forecasts is also considered. In section 3.1 I found that the variable selection algorithm used to identify the LMESS model at each estimation period does not have 100% accuracy. To investigate whether wrongly selected models affect the forecast accuracy, I also test the most common linear mixed-effect state-space model (MCLME) forecasts. For each dependent variable, I use the most commonly selected model specification across the ten estimation periods. The assumption is thus imposed that the optimal LMESS model specification does not change over time.

The multivariate linear regression (MLR) forecasts were constructed by using a backward selection procedure. Starting with all covariates, I estimate the model and exclude the covariate with the highest p-value which is insignificant at the 90% confidence level. If all covariates in the model are significant, I proceed by estimating the least squares coefficient vector β. The forecasts are calculated with the realised covariate values in the forecast period.

The benchmark forecasts more elaborate forecasting methods want to improve upon is given by the long-term mean. When working with climate data, this long-term mean is often referred to as the climatology forecast. In this study, I calculate the average value of each dependent variable as follows. First, I split each variable in twelve distinct time series, each corresponding to a different month of the year. For each month, I calculate the average of the time series, which gives twelve values. Over the forecast period, the forecast for each year is given by these twelve values.

5.1 Selected Variables

Table 4 shows which covariates are included in the most common linear mixed-effect state-space model and the multivariate linear regression model. For both models, different measurement sites show different dynamics. For the MCLME model, the inclusion of covariates as fixed and random does not coincide between any of the five measurement sites. However, the manner in which variables are included does show structure. For example, all fixed effects that are included at any measurement site are significant at the 95% level. The intercept is always a fixed effect if it is included, except for logit(p_{rain}) in De Kooy, where it is random. The level of CO_2 is always included as a random effect for all dependent variables except one case. Furthermore, the state-space coefficients tend to be negative for the temperature time series, although this is not significant. For the rain variable time series, the CO_2 level is significantly negative in both De Kooy and Vlissingen and is positively and negatively significant for $\log(\mu_{rain})$ in Beek and Eelde, respectively. The NAO index is only significant for both temperature time series at all measurement sites and is always significantly positive, not only if it is included as fixed, but also if the NAO index is a random effect. If the cyclone density ρ_{cyc} and anti-cyclone density $\rho_{anti\text{-}cyc}$ are included, they are always included fixed effects. Furthermore, if both are included in a model, which is the case in De Bilt, De Kooy, Eelde, and Beek, they have opposite sign and are of roughly equal size. The relative humidity RH is included as a random effect in all dependent time series at all measurement sites except for the $\log(\mu_{rain})$ in De Bilt, Eelde, and Beek. The state-space terms tend to be negative, but the coefficients are only significantly negative for the rain variable time series in De Bilt, De Kooy, Vlissingen, and Beek.

Table 4

Variables Included in LMESS and MLR

This table reports the coefficients that are used for forecasting for the dependent variables at all locations for the longest estimation period between January 1959 and December 2007. In column three to eight, the coefficients for the Most Common linear mixed-effect state-space model is given. Empty cells indicate the covariate is not included. Covariates included as random are indicated by + or −, which indicates the sign of the majority of smoothed state coefficients. Fixed effect coefficients are given as the LMESS model estimate. Columns nine to twelve report the coefficients as found by initial general multivariate linear regression. Green cells indicate the covariates included in the final model used for forecasting.

		Most Common LMESS model						Multivariate Linear Regression					
Location	Variable	Intercept	CO_2	NAO	ρ_{cyc}	$\rho_{anti\text{-}cyc}$	RH	Intercept	CO_2	NAO	ρ_{cyc}	$\rho_{anti\text{-}cyc}$	RH
De Bilt	$\text{logit}(p_{rain})$	-1.20^{*}			0.18^{*}		$-^{*}$	-5.85^{**}	0.00	-0.01	0.21^{**}	-0.01	5.07^{**}
	$\log(\mu_{rain})$	5.00^{*}	$-$		0.27^{*}			2.24	0.01	-0.09	0.28^{**}	0.02	0.85
	$\Delta_{12}Temp_{max}$	0.12^{*}	$-$	0.56^{*}			$-$	5.90^{**}	0.00	0.62^{**}	0.04	-0.12	-6.21^{**}
	$\Delta_{12}Temp_{min}$		$-$	$+^{*}$	0.29^{*}	-0.25^{*}	$-$	-0.85	0.00	0.40^{**}	0.34^{**}	-0.22	0.32
De Kooy	$\text{logit}(p_{rain})$	$-^{*}$	$-^{*}$		0.21^{*}	-0.36^{*}	$-^{*}$	-6.32^{**}	0.00	-0.05^{**}	0.30^{**}	-0.12^{**}	5.91^{**}
	$\log(\mu_{rain})$	4.63^{*}	$-^{*}$		0.26^{*}		$-^{*}$	7.70^{**}	0.00	-0.16^{**}	0.23^{**}	0.08	-3.90^{**}
	$\Delta_{12}Temp_{max}$		$-$	0.48^{*}			$-$	2.98	0.00	0.55^{**}	0.10	-0.11	-3.51
	$\Delta_{12}Temp_{min}$		$-$	0.39^{*}			$-$	0.53	0.00	0.47^{**}	0.26^{**}	-0.21	-0.93
Eelde	$\text{logit}(p_{rain})$	-1.10^{*}	$+$		0.17^{*}		$+$	-5.53^{**}	0.00	0.00	0.22^{**}	-0.04	5.04^{**}
	$\log(\mu_{rain})$		$+^{*}$					5.87^{**}	0.00	-0.06	0.16	0.11	-2.53^{**}
	$\Delta_{12}Temp_{max}$		$-$	0.57^{*}			$+$	5.71^{**}	0.00	0.67^{**}	0.12	-0.16	-6.13^{**}
	$\Delta_{12}Temp_{min}$		$-$	$+^{*}$	0.41^{*}	-0.38^{*}	$-$	-1.15	0.00	0.45^{**}	0.37^{**}	-0.30^{**}	0.72
Vlissingen	$\text{logit}(p_{rain})$		$-^{*}$		-0.16^{*}		$-^{*}$	-6.13^{**}	0.00	-0.05^{**}	0.22^{**}	-0.03	5.24^{**}
	$\log(\mu_{rain})$	5.48^{*}	$-^{*}$				$-^{*}$	6.71^{**}	0.00	-0.20^{**}	0.10	0.15	-2.22
	$\Delta_{12}Temp_{max}$		$-$	0.48^{*}			$-$	4.53^{**}	0.00	0.55^{**}	0.08	-0.15	-5.05^{**}
	$\Delta_{12}Temp_{min}$		$-$	0.33^{*}			$+$	-0.20	0.00	0.44^{**}	0.23^{**}	-0.19	-0.10
Beek	$\text{logit}(p_{rain})$	-1.27^{*}	$+$		0.19^{*}		$-^{*}$	-5.13^{**}	0.00	-0.02	0.21^{**}	-0.01	4.21^{**}
	$\log(\mu_{rain})$	5.03^{*}	$-^{*}$			0.21^{*}		4.95^{**}	0.00	-0.06	-0.15	0.43^{**}	-0.33
	$\Delta_{12}Temp_{max}$	0.10^{*}	$-$	0.53^{*}			$-$	9.02^{**}	0.00	0.58^{**}	0.10	-0.22	-9.39^{**}
	$\Delta_{12}Temp_{min}$		$-$	$+^{*}$	0.09^{*}	-0.06^{*}	$-$	0.80	0.00	0.37^{**}	0.34^{**}	-0.19	-1.72

* Estimate is statistically significant at the 5%, or for state-space terms the sign is positive or negative in a significantly non-random number of cases.

** Linear regression coefficient is statistically significant at the 5% level.

For the multivariate linear regression model, table 4 shows both the results of a an initial regression with all covariates as well as the final model after the general-to-specific selection procedure. In the initial regression, the significant coefficient across different measurement sites have different values. Also, the final multivariate linear regression models used for forecasting, which are indicated by the coloured cells, show different patterns at all five measurement sites. This affirms the finding of the MCLME model that all measurement sites have different dynamics. An example is the NAO index, which is significant for all four dependent time series in the initial regression in De Kooy and Vlissingen, but only significant for the temperature time series in De Bilt, Eelde, and Beek. The NAO index in De Kooy and Vlissingen is also the only variable to not be included in the final multivariate linear regression forecast model if it was significant in the general regression. In all other cases, the 95% significance is a reliable indicator that the variable is important to forecasting. However, variables not significant at first are included in the final model. The anti-cyclone density $\rho_{anti\text{-}cyc}$ is only significant in the general regression for three out of 20 cases, but is included in forecasting in nine cases. The level of atmospheric CO_2 is never significant in the initial regression and is included only once in the final forecasts, namely for $\log(\mu_{rain})$ in Eelde. This is probably due to the strong trending and seasonal behaviour of the CO_2 level time series, given that the dependent time series do not show these characteristics as strongly.

Table 4 also shows that the most common linear mixed-effect state-space and multivariate linear regression models use roughly the same covariates in De Bilt and Beek. In De Bilt, the only differences are in the inclusion of CO_2 level and relative humidity RH for the minimum temperature series. For the other three measurement sites, the NAO index also follows the same pattern of inclusion in both models. The most notable exception is the CO_2 level. This covariate is only once included as significant covariate in the multivariate linear regression model, whereas it is included as random effect in the most common linear mixed-effect state-space model in all but one case. Relative humidity, on the other hand, is included as covariate for a majority of dependent time series in both the LMESS and MLR model, apart from the minimum temperature. So although it is never remarked as a fixed effect by the linear mixed-effect state-space model, it can still be included as a covariate in linear regression, where the coefficient is fixed.

5.2 Forecasting Results

Table 5 reports the average percentage reduction in root mean squared forecast error relative to the climatology forecasts for the LMESS and MLR model. For both temperature data series, both the LMESS and MLR model provide slightly better forecasts than climatology in most cases across all five measurement locations. Comparing both models, all values are within 5% of equal RMSFE, which leads to the conclusion that the models are indistinguishable in terms of forecast improvement for temperature. In the case of rainfall, a more pronounced difference between LMESS and MLR can be seen. Neither method can improve on the long term mean to forecast the average rainfall on rainy days, with RMSFE increasing by at least 5%. In Eelde, the RMSFE for the LMESS forecasts is more than double that of the climatology forecasts. The reason for this inaccuracy probably lies in the chaotic and complex nature of how rain clouds develop and the moment at which it rains. Also, there can be high deviations in measured rainfall across only a small area of land, whereas the covariates considered in this study have a more large scale character. The MLR model does provide better forecasts than the LMESS model for $\log(\mu_{rain})$ in all cases. For the number of rainy days, large differences are reported between different weather stations. In De Bilt, LMESS and MLR both reduce the RMSFE significantly, where the MLR model shows a higher reduction than LMESS. At De Kooy, Eelde, Vlissingen, and Beek the multivariate regression forecasts show similar reductions in RMSFE compared to the long-term mean forecasts as at De Bilt. The RMSFE of the LMESS model at those four stations for $\text{logit}(p_{rain})$ is not significantly better than the long-term mean RMSFE. The forecasting performance of the LMESS model at De Kooy and Vlissingen is even significantly worse than climatology.

Furthermore, table 5 shows significant differences between the RMSFE reduction of the linear mixed-effect state-space model and the RMSFE reduction of the MCLME approach, where the same model is assumed for all estimation periods. Apart from the proportion of rainy days in De Kooy and Vlissingen and mean rainfall in Eelde, the MCLME model shows a similar or higher RMSFE reduction than the non-restricted LMESS model. The main differences are seen in De Bilt, De Kooy, Vlissingen, and Beek, where a large increase in RMSFE compared to climatology is observed from the LMESS model when forecasting $\text{logit}(p_{rain})$. Imposing the most common model assumption improves the RMSFE reduction value. In Eelde, the RMSFE for $\log(\mu_{rain})$ is more than double the

Table 5

RMSFE reduction results

This table shows the percentage RMSFE reduction in predictive accuracy compared to the long term mean forecast for the linear mixed-effect state-space model (LMESS), the most common LMESS model (MCLME), and multivariate linear regression model (MLR) for three forecast horizons. The reported value is an average over ten forecast periods with the same length, where different forecasts are based on an expanding window estimation period. Red and green cell colours indicate a RMSFE reduction being smaller than -3% or bigger than 3% respectively, with darker shading for larger reduction absolute values.

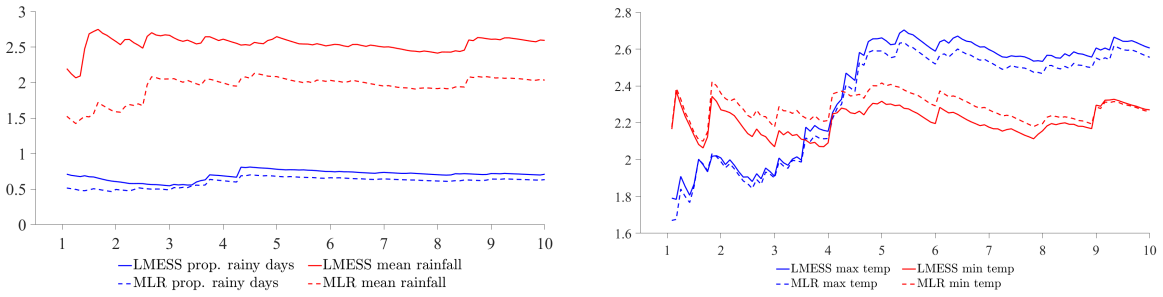
	1 year			5 year			10 years		
Variable	LMESS	MCLME	MLR	LMESS	MCLME	MLR	LMESS	MCLME	MLR
De Bilt									
$\text{logit}(p_{rain})$	6.86	5.28	3.77	6.42	5.68	17.52	5.18	4.49	15.49
$\log(\mu_{rain})$	−26.36	−4.68	−5.13	−19.60	−4.68	−4.62	−16.64	−5.17	−5.02
$\Delta_{12}Temp_{max}$	−1.53	0.88	2.85	0.90	1.40	3.49	1.43	2.49	4.24
$\Delta_{12}Temp_{min}$	0.61	0.61	−0.59	0.00	−0.03	−0.26	0.10	0.02	0.60
De Kooy									
$\text{logit}(p_{rain})$	−7.28	−12.51	1.76	−5.26	−12.41	4.77	−10.79	−14.82	4.60
$\log(\mu_{rain})$	−30.91	−9.21	−9.82	−27.48	−6.14	−6.66	−34.61	−8.53	−8.66
$\Delta_{12}Temp_{max}$	1.33	1.53	0.91	1.51	1.60	1.35	2.12	2.19	2.05
$\Delta_{12}Temp_{min}$	0.79	2.02	1.57	1.30	1.46	1.34	1.21	1.67	1.23
Eelde									
$\text{logit}(p_{rain})$	−0.88	−1.08	8.42	0.76	0.47	10.02	2.12	1.74	8.90
$\log(\mu_{rain})$	−192.87	−235.40	−5.86	−155.11	−194.96	−6.13	−150.50	−191.98	−9.77
$\Delta_{12}Temp_{max}$	1.87	1.55	2.82	1.65	2.08	3.76	1.94	3.06	4.66
$\Delta_{12}Temp_{min}$	1.63	2.04	1.83	0.75	1.04	2.18	0.45	0.13	2.09
Vlissingen									
$\text{logit}(p_{rain})$	−5.97	−14.09	5.76	−6.02	−18.30	14.63	−7.35	−18.67	14.08
$\log(\mu_{rain})$	−33.73	−12.95	−12.35	−24.31	−9.81	−8.06	−28.27	−11.20	−9.26
$\Delta_{12}Temp_{max}$	1.28	0.61	0.86	1.09	1.02	1.74	1.95	2.16	2.97
$\Delta_{12}Temp_{min}$	0.60	0.75	−0.08	0.17	0.68	0.30	0.44	1.73	1.61
Beek									
$\text{logit}(p_{rain})$	−1.61	−0.03	11.87	−0.68	6.30	18.11	0.17	7.31	18.73
$\log(\mu_{rain})$	−19.98	−7.42	−7.28	−15.40	−8.73	−8.67	−13.41	−8.25	−7.91
$\Delta_{12}Temp_{max}$	−0.03	−0.01	3.81	0.54	0.51	3.80	1.94	1.93	4.99
$\Delta_{12}Temp_{min}$	−0.12	1.10	−1.16	0.23	0.21	−0.57	0.26	0.28	0.55

climatology RMSFE for both for the LMESS model and MCLME model. This increase in forecast error is most likely due to the limited number of included covariates. Table 4 shows that for this time series, the most common model only includes the CO_2 level as a random effect. These findings indicate the LMESS model is susceptible to wrong model

specifications. If covariates are wrongly identified as fixed or random effects, or excluded when they should not be, the root mean squared forecast error grows large compared to climatalogy, influencing the average forecast accuracy.

Table 5 also shows that the RMSFE reduction compared to the long-term mean forecast increases for larger forecast periods in most cases. This can be the result of either a changing climate, where the estimation period average is no longer representative at 10 years beyond the estimation period. But it can also be caused by the assumption of perfect foresight in the forecast, where realised covariate data was used to calculate the forecast. The last assumption eliminates a source of forecast noise which can not be eliminated in real life applications.

Figures 6a and 6b show how the root mean squared forecast error behaves during a ten-year forecast period between January 2003 and December 2012 for the linear mixed-effect state-space and multivariate linear regression models. The figures support the findings in Table 5, as the RMSFE of the LMESS forecasts is consistently larger than the equivalent MLR forecasts for the rain related forecasts and the maximum temperature times series. Only for minimum temperature time series, does the LMESS model notably outperform the MLR model. For all four dependent variables, the behaviour of the RMSFE as a function of forecast horizon is very similar between both considered models. For the proportion of rainy days and the minimum temperature, the RMSFE only moves within a small window, and is close to constant between five- and ten-year forecast horizons. The RMSFE for mean rainfall does show more variability in the shorter forecast period,

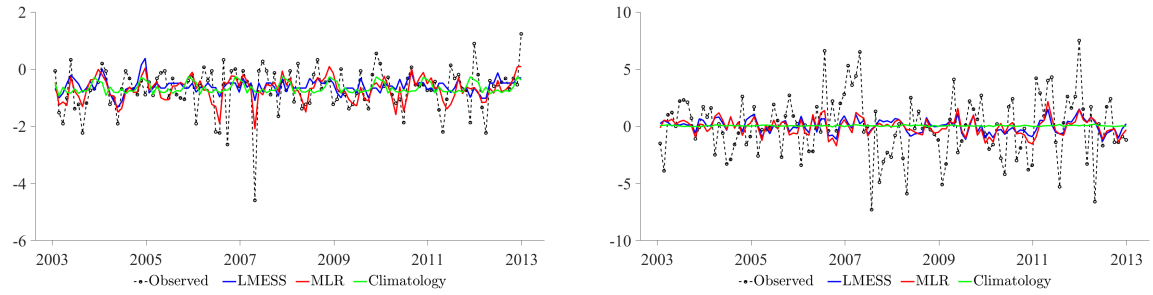


(a) RMSFE of forecasts rainfall **(b)** RMSFE of forecasts temperature

Figure 6. This figure shows the comparison between the LMESS and MLR model RMSFE as a function of forecast horizon for the weather station in De Bilt between January 2003 and December 2012. The forecast period spans between January 1959 and December 2002. The two figures show graphs of both models for (a) the logit proportion of rainy days and log mean rainfall and (b) 12-month difference in maximum and minimum temperature.

but stabilises after three years, after which it remains close to constant. The maximum temperature forecast RMSFE shows an upward slope between one- and five-year forecasts. This increase in RMSFE for larger forecast periods is in line with normal behaviour of RMSFE as a function of forecast horizon. Both lines do however flatten after the 5-year forecasts and decrease for longer forecast horizons.

A comparison between forecasts by the linear mixed-effect state-space model, multivariate linear regression, climatology, and realised values is shown in figures 7a and 7b for two dependent variables in De Bilt. The observed data is shown for the period between January 2003 and December 2012 together with five year forecasts. So, for 2003, the forecasts are based on the estimation period ending in 1998. For the proportion of rainy days in figure 7a, both the LMESS and MLR show better ability to capture the real data dynamics that climatology, which supports the findings in table 5. In general, the MLR forecasts do more closely resemble the realised data, but the LMESS forecasts follow the dynamics of the MLR forecasts closely across the ten years. As expected, the climatology forecasts do not reflect the realised data as well as the two models using covariate data.



(a) Five year forecast logit(p_{rain})

(b) Five year forecast $\Delta_{12}Temp_{max}$

Figure 7. This figure shows the accuracy of 5-year forecasts made using the linear mixed-effect state-space model, multivariate linear regression, and climatology compared to the realised values at De Bilt in the period between January 2003 and December 2012. The two dependent variables are (a) logit proportion of rainy days per month and (b) twelve month difference in average maximum temperature.

The time series forecast for maximum temperature in figure 7b shows an even closer match between the LMESS and MLR forecasts. Both show more dynamics than the climatology forecasts, but fail to forecast the dynamics of the yearly temperature change. As climatology forecasts only seem to predict a mean close to zero, one would expect the more elaborate models like LMESS and MLR to improve forecast accuracy, especially considering the assumed perfect foresight on covariates. But figure 7b supports the results

in table 5 that the forecasts are improved, but not with a significant percentage reduction in RMSFE. These findings show that the covariates included in this study do not span all explanatory factors on changing temperature in the Netherlands, which is a more random process than can be captured by any of the models considered in this study.

6 Conclusion

The study of Kokic et al. (2011) was a first step to a new way of forecasting climate variables. They identified two possibilities of further research that are addressed in this thesis. Firstly, I build on the work of Kokic et al. (2011) to propose an algorithm that identifies explanatory covariates to use in their linear mixed-effect state-space model formulation. And secondly, I explore the ability of the LMESS model to forecast in monthly average climate data in the Netherlands.

By means of a simulation study I find that the algorithm originally used by Kokic et al. (2011) to identify covariates as fixed and random effects is suboptimal and leads to wrong model specifications in at least 80% of cases. To improve selection accuracy, I propose a new selection algorithm based on a state-space formulation with fixed parameters as first described by Chow (1984). This new algorithm significantly improves the variable selection accuracy to at least 50% if enough observations are present and over 70% under certain data assumptions. I find that the variable selection algorithm proposed by Bondell et al. (2010) does not consistently achieve its original accuracy when the observation covariance is adjusted for serial correlation imposed by the linear mixed-effect state-space model's state equation. Although it identifies the correct model in 76% of simulations for one case, it fails to select the correct model even once in other scenarios.

In a second simulation study, I demonstrate the importance of selection of the correct model on forecast accuracy. I find that the forecasts of a correct model specification are significantly better than forecasts based on wrong model specifications. Out of the considered wrong specifications, none could outperform the true model at short forecast horizons and only one wrong specification could improve significantly on the forecasts by the true model at the longest horizon, but not consistently across two dependent variables.

Applying the new methods to climate data in the Netherlands, I find that the linear mixed-effect state-space model does not consistently give better forecasts than a multivariate linear regression model or climatology. For changes in temperature, all three methods yield similar results in terms of root mean forecast squared error, all within 5% of one another across horizons of one, five, and ten years. For all measurement sites considered in this study, not once are the linear mixed-effect state-space model and multivariate linear regression model able to improve on climatology forecasts for the average rain on rainy days. For the proportion of rainy days, I find the linear mixed-effect state-space

forecast to be sensitive to wrong model specification, only significantly outperforming the climatology benchmark in De Bilt.

The models presented in this research can be build upon by further research to improve applicability on real climate forecasts. Although the variable selection method based on the state-space formulation by Chow (1984) does improve model selection accuracy, it is not perfect. The algorithm proposed by Bondell et al. (2010) should increase the identification accuracy if the longitudinal covariance matrix can be identified correctly. However, the state estimation algorithms used in this study might not be sufficient to address this issue. The influence of covariate prediction uncertainty on the forecast accuracy by the linear mixed-effect state-space model also needs to be addressed before definitive conclusions can be drawn on applicability. Forecasts across different methods may be more accurately compared by the general matrix of the forecast-error second-moment approach proposed by Hendry & Martinez (2017). Their method yields an invariant measure of forecast accuracy that results in more robust conclusions on model preference than the average root mean squared forecast error reduction used in this study.

References

Andersen, T. G. & Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International economic review*, 885–905.

Arrhenius, S. (1896). On the influence of carbonic acid in the air upon the temperature of the ground. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(251), 237–276.

Beckmann, B.-R. & Buishand, A. T. (2002). Statistical downscaling relationships for precipitation in the Netherlands and North Germany. *International Journal of Climatology*, *22*(1), 15–32.

Beniston, M. (2004). The 2003 heat wave in Europe: A shape of things to come? An analysis based on Swiss climatological data and model simulations. *Geophysical Research Letters*, *31*(2).

Beniston, M., Stephenson, D. B., Christensen, O. B., Ferro, C. A. T., Frei, C., Goyette, S., ... Woth, K. (2007). Future extreme events in European climate: an exploration of regional climate model projections. *Climatic Change*, *81*(1), 71–95.

Bondell, H. D., Krishna, A., & Ghosh, S. K. (2010). Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models. *Biometrics*, *66*(4), 1069–1077.

Calanca, P., Bolius, D., Weigel, A. P., & Liniger, M. A. (2011). Application of long-range weather forecasts to agricultural decision problems in Europe. *The Journal of Agricultural Science*, *149*(1), 15–22.

Cassou, C., Terray, L., & Phillips, A. S. (2005). Tropical Atlantic Influence on European Heat Waves. *Journal of Climate*, *18*(15), 2805–2811.

Chen, Z. & Dunson, D. B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, *59*(4), 762–769.

Chow, G. C. (1984). Random and changing coefficient models. *Handbook of econometrics*, *2*, 1213–1245.

Convery, F. J. & Wagner, G. (2015). Reflections–managing uncertain climates: some guidance for policy makers and researchers. *Review of Environmental Economics and Policy*, *9*(2), 304–320.

Donat, M. G., Leckebusch, G. C., Pinto, J. G., & Ulbrich, U. (2009). Examination of wind storms over Central Europe with respect to circulation weather types and NAO phases. *International Journal of Climatology*, *30*(9), 1289–1300.

Dorland, C., Tol, R. S., & Palutikof, J. P. (1999). Vulnerability of the Netherlands and Northwest Europe to storm damage under climate change. *Climatic change*, *43*(3), 513–535.

Durbin, J. & Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology*, *27*(12), 1547–1578.

Haarsma, R. J., Hazeleger, W., Severijns, C., Vries, H., Sterl, A., Bintanja, R., ... Brink, H. W. (2013). More hurricanes to hit western Europe due to global warming. *Geophysical Research Letters*, *40*(9), 1783–1788.

Heij, C., De Boer, P., Franses, P. H., Kloek, T., & Van Dijk, H. K. (2004). *Econometric methods with applications in business and economics*. OUP Oxford.

Hendry, D. F. & Martinez, A. B. (2017). Evaluating multi-step system forecasts with relatively few forecast-error observations. *International Journal of Forecasting*, *33*(2), 359–372.

Hoerling, M. P., Hurrell, J. W., & Xu, T. (2001). Tropical Origins for Recent North Atlantic Climate Change. *Science*, *292*(5514), 90–92.

Hurrell, J. W. (1995). Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science*, *269*(5224), 676–679.

Kerr, R. A. (1997). A New Driver for the Atlantic's Moods and Europe's Weather? *Science*, *275*(5301), 754–755.

Kokic, P., Crimp, S., & Howden, M. (2011). Forecasting climate variables using a mixed-effect state-space model. *Environmetrics*, *22*(3), 409–419.

Koninklijk Nederlands Meteorologisch Instituut (KNMI). (2018). Maand- en jaarwaarden. Retrieved from http://www.knmi.nl/nederland-nu/klimatologie/maandgegevens.

Kukkonen, J., Olsson, T., Schultz, D. M., Baklanov, A., Klein, T., Miranda, A. I., ... Eben, K. (2012). A review of operational, regional-scale, chemical weather forecasting models in Europe. *Atmospheric Chemistry and Physics*, *12*(1), 1–87.

National Weather Service Climate Prediction Center. (2018). Monthly Teleconnection Index: North Atlantic Oscillation (NAO). Retrieved from http://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/nao.shtml.

Neiburger, M., Edinger, J. G., & Bonner, W. D. (1971). Fronts and Cyclones. In *Understanding Our Atmospheric Environment* (Chap. 11, pp. 249–273). W. H. Freemand and Company.

Nelson, D. B. & Foster, D. P. (1995). Filtering and forecasting with misspecified ARCH models II: making the right forecast with the wrong model. *Journal of Econometrics*, *67*(2), 303–335.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., ... Dasgupta, P., et al. (2014). *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change*. IPCC.

Perry, J. H. (1950). *Chemical engineers' handbook*. ACS Publications.

Pourahmadi, M. & Daniels, M. (2002). Dynamic conditionally linear mixed models for longitudinal data. *Biometrics*, *58*(1), 225–231.

Schaap, M., Timmermans, R. M., Roemer, M., Boersen, G., Builtjes, P., Sauter, F., ... Beck, J. (2008). The LOTOS-EUROS model: description, validation and latest developments. *International Journal of Environment and Pollution*, *32*(2), 270–290.

Shumway, R. H. & Stoffer, D. S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *Journal of time series analysis*, *3*(4), 253–264.

Swanson, N. R. & White, H. (1997). A model selection approach to real-time macroeconomic forecasting using linear models and artificial neural networks. *Review of Economics and Statistics*, *79*(4), 540–550.

Tank, A. M. G. K., Zwiers, F. W., & Zhang, X. (2009). Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation. *World Meteorological Organization*.

Tans, P. (NOAA/ESRL) & Keeling, R. (Scripps Institution of Oceanography). (2018). Mauna Loa CO_2 monthly mean data. Retrieved from https://www.esrl.noaa.gov/gmd/ccgg/trends/data.html.

Wackerly, D., Mendenhall, W., & Scheaffer, R. L. (2014). *Mathematical statistics with applications*. Cengage Learning.

Wetherald, R. T. & Manabe, S. (1995). The Mechanisms of Summer Dryness Induced by Greenhouse Warming. *Journal of Climate*, *8*(12), 3096–3108.

Appendices

A EM-algorithm Chow

In this section I outline the E- and M-step of the algorithm used for parameter estimation in the mixed-effect model formulation of Chow (1984) in equations (9) and (10). Firstly, the E-step is used to find estimates of the underlying states and state covariance conditional on all observations. Despite the extensions to the state vector, there is no change in the filter and smoother equations compared to the linear state-space model equations. Thus, the E-step is given by the following equations.

$$\begin{bmatrix} \beta_{i|t} \\ \alpha_{ijt|t} \end{bmatrix} = B_{ij} \begin{bmatrix} \beta_{i|t-1} \\ \alpha_{ijt-1|t-1} \end{bmatrix} + K_t^* e_t^* \qquad V_t^* = B_{ij} \hat{\Sigma}_{t-1|t-1}^* B_{ij}' + \begin{bmatrix} 0 & 0 \\ 0 & Q_{ij} \end{bmatrix}$$

$$\hat{\Sigma}_{t|t}^* = V_t^* - K_t^* X_{ijt}^* V_t^* \qquad K_t^* = V_t X_{ijt}^{*\prime} (R_{ij} + X_{ijt}^* V_t X_{ijt}^{*\prime})^{-1}$$

$$e_t^* = y_{ijt} - X_{ijt}^* B_{ij} \begin{bmatrix} \beta_{i|t-1} \\ \alpha_{ijt-1|t-1} \end{bmatrix}$$

Using these estimates, the smoothed estimates based on the full data set of all observations. There is no difference between these estimates and the standard linear state-space model. Thus the smoothed estimates for the smoothed states, smoothed covariance and transition covariance are given by

$$\begin{bmatrix} \beta_{i|T} \\ \alpha_{ijt-1|T} \end{bmatrix} = \begin{bmatrix} \beta_{i|T} \\ \alpha_{ijt|T} \end{bmatrix} + \hat{\Sigma}_{t-1|t-1}^* B_{ij}' V_t^{*-1} \left(\begin{bmatrix} \beta_{i|T} \\ \alpha_{ijt|T} \end{bmatrix} - \begin{bmatrix} \beta_{i|t} \\ \alpha_{ijt|t} \end{bmatrix} \right)$$

$$\hat{\Sigma}_{t-1|T}^* = \hat{\Sigma}_{t-1|t-1}^* - \hat{\Sigma}_{t-1|t-1}^* B_{ij}' V_t^{*-1} (V_t^* - \hat{\Sigma}_{t|T}^*) V_t^{*-1} B_{ij} \hat{\Sigma}_{t-1|t-1}^* + \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{ijt-1|T} \alpha_{ijt-1|T}' \end{bmatrix}$$

$$\hat{\Sigma}_{t,t-1|T}^* = \hat{\Sigma}_{t|T}^* V_t^{*-1} B_{ij} \hat{\Sigma}_{t-1|t-1}^* + \begin{bmatrix} 0 & 0 \\ 0 & \alpha_{ijt|T} \alpha_{ijt-1|T}' \end{bmatrix}$$

With these estimated matrices, I can now find the maximum likelihood estimates for this model. The M-step is equivalent to the M-step in the standard linear state-space model,

with some restrictions imposed by the inclusion of constant parameters.

$$\hat{B}_{ij} = \begin{bmatrix} I & 0 \\ 0 & \left(\left(\sum_{t=2}^{T} \hat{\Sigma}^*_{t,t-1|T} \right) \left(\sum_{t=2}^{T} \hat{\Sigma}^*_{t-1|T} \right)^{-1} \right)_{(q_i+1,2q_i;q_i+1,2q_i)} \end{bmatrix}$$

$$\hat{R}_{ij} = T^{-1} \sum_{t=1}^{T} \left(y_t - X^*_{ijt} \begin{bmatrix} \beta_{i|T} \\ \alpha_{ijt|T} \end{bmatrix} \right) \left(y_t - X^*_{ijt} \begin{bmatrix} \beta_{i|T} \\ \alpha_{ijt|T} \end{bmatrix} \right)'$$

$$\hat{Q}_{ij} = \left((T-1)^{-1} \sum_{t=2}^{T} \left[\hat{\Sigma}^*_{t|T} - \hat{B}_{ij} \hat{\Sigma}^{*\prime}_{t,t-1|T} - \hat{\Sigma}^*_{t,t-1|T} \hat{B}'_{ij} + \hat{B}_{ij} \hat{\Sigma}^*_{t-1|T} \hat{B}'_{ij} \right] \right)_{(q_i+1,2q_i;q_i+1,2q_i)}$$

The most important restriction is to set the top left of $\hat{B}_{ij}$ as identity matrix.

B Bondell's M-step

This section is based on web appendix B from the paper of Bondell et al. (2010). For notation convenience, we remove the subscript i, as the procedure is the same for all dependent variables. First, based on the likelihood function in Equation (15), we define $(y^*, X^*, Z^*) = (\tilde{\Omega}^{-1/2}y, \tilde{\Omega}^{-1/2}X, \tilde{\Omega}^{-1/2}Z)$. At each iteration we first update three additional variables related to the distribution of the mean random effects. Given the likelihood function in Equation (14), the conditional distribution $\bar{\alpha}|y, \phi \sim N(\hat{\bar{\alpha}}, U)$, with

$$\hat{\bar{\alpha}}^{(\omega)} = \big(\tilde{\Gamma}'^{(\omega)}\tilde{D}^{(\omega)}Z^{*\prime}Z_i^*\tilde{D}^{(\omega)}\tilde{\Gamma}^{(\omega)} + I_{m\cdot q}\big)^{-1}\big(Z^*\tilde{D}^{(\omega)}\tilde{\Gamma}^{(\omega)}\big)'\big(y^* - X^*\beta^{(\omega)}\big), \tag{23}$$

$$\text{and} \quad U^{(\omega)} = \sigma^{2(\omega)}\big(\tilde{\Gamma}'^{(\omega)}\tilde{D}^{(\omega)}Z^{*\prime}Z^*\tilde{D}^{(\omega)}\tilde{\Gamma}^{(\omega)} + I_{m\cdot q}\big)^{-1}, \tag{24}$$

where ω indicates the iterations of the EM algorithm. The starting values for $\omega = 0$ are set as the restricted maximum likelihood estimates for the model in Equation (11). The estimate for $\sigma^{2(\omega)}$ at iteration ω is updated as

$$\sigma^{2(\omega)} = \big(y^* - X^*\beta^{(\omega)}\big)'\big(Z^*\tilde{D}^{(\omega)}\tilde{\Gamma}^{(\omega)}\tilde{\Gamma}'^{(\omega)}\tilde{D}^{(\omega)}Z^{*\prime} + I_{m\cdot T}\big)^{-1}\big(y^* - X^*\beta^{(\omega)}\big)/(m\cdot T). \tag{25}$$

We regard these three variables as constant when performing the optimisation over ϕ. Omitting terms that do not involve the variable ϕ, we can rewrite the estimation step in Equation (15) as

$$\begin{aligned} &g(\beta, d|\phi^{(\omega)}) = \\ &\begin{bmatrix}\beta\\ d\end{bmatrix}' \begin{bmatrix} X^{*\prime}X^* & X^{*\prime}Z^*\mathrm{diag}(\tilde{\Gamma}\hat{\bar{\alpha}}^{(\omega)})(\mathbf{1}_m \otimes I_q) \\ (\mathbf{1}_m \otimes I_q)'\mathrm{diag}(\tilde{\Gamma}\hat{\bar{\alpha}}^{(\omega)})Z^{*\prime}X^* & (\mathbf{1}_m \otimes I_q)'\big(W \bullet \tilde{\Gamma}\tilde{G}^{(\omega)}\tilde{\Gamma}'\big)(\mathbf{1}_m \otimes I_q) \end{bmatrix} \begin{bmatrix}\beta\\ d\end{bmatrix} \\ &- 2y^{*\prime}\begin{bmatrix} X^* & Z^*\mathrm{diag}(\tilde{\Gamma}\hat{\bar{\alpha}}^{(\omega)})(\mathbf{1}_m \otimes I_q)\end{bmatrix}\begin{bmatrix}\beta\\ d\end{bmatrix} + \lambda_m\Big(\sum_{n=1}^{p}\frac{|\beta_n|}{|\bar{\beta}_n|} + \sum_{n=1}^{q}\frac{|d_n|}{|\bar{d}_n|}\Big), \end{aligned} \tag{26}$$

where $\bullet$ is the Hadamard product operator, $W = Z^{*\prime}Z^*$ a symmetric block matrix, and $\tilde{G}^{(\omega)} = U^{(\omega)} + \hat{\bar{\alpha}}^{(\omega)}\hat{\bar{\alpha}}^{(\omega)\prime}$. However, due to the absolute element-wise sum in the penalty term, this is a non-standard optimisation problem. If we write $\beta = \beta^+ - \beta^-$, where β^+ and β^- are non-negative and only one non-zero, and $|\beta| = \beta^+ + \beta^-$, the optimisation becomes a quadratic programming problem.

$$\text{minimise} \quad \begin{bmatrix} \beta^+ \\ \beta^- \\ d \end{bmatrix}' \begin{bmatrix} X_\dagger^{*\prime} X_\dagger^* & X_\dagger^{*\prime} Z^* \text{diag}(\tilde{\Gamma}\hat{\tilde{\alpha}}^{(\omega)})(\mathbf{1}_m \otimes I_q) \\ (\mathbf{1}_m \otimes I_q)' \text{diag}(\tilde{\Gamma}\hat{\tilde{\alpha}}^{(\omega)}) Z^{*\prime} X_\dagger^* & (\mathbf{1}_m \otimes I_q)' \big(W \bullet \tilde{\Gamma}\tilde{G}^{(\omega)}\tilde{\Gamma}'\big)(\mathbf{1}_m \otimes I_q) \end{bmatrix} \begin{bmatrix} \beta^+ \\ \beta^- \\ d \end{bmatrix}$$

$$- 2\Bigg(y^{*\prime} \Big[X_\dagger^* \quad Z^* \text{diag}(\tilde{\Gamma}\hat{\tilde{\alpha}}^{(\omega)})(\mathbf{1}_m \otimes I_q) \Big] + \lambda_m \Big[\tfrac{1}{|\beta_1|}, \cdots, \tfrac{1}{|\beta_p|}, \tfrac{1}{|\beta_1|}, \cdots, \tfrac{1}{|\beta_p|}, \tfrac{1}{|d_1|}, \cdots, \tfrac{1}{|d_q|} \Big] \Bigg) \begin{bmatrix} \beta^+ \\ \beta^- \\ d \end{bmatrix} \tag{27}$$

$$\text{subject to:} \quad \beta^+ \geq 0, \quad \beta^- \geq 0, \quad d \geq 0,$$

where the matrix $X_\dagger^* = [X^* \ -X^*]$.

With a solution for $(\beta', d')'$, the new optimal value for γ has a closed form solution given by

$$\gamma_* = \big(P^{(\omega)}\big)^- \big[R'^{(\omega)}(y^* - X^*\beta) - T^{(\omega)}\big], \tag{28}$$

where $P^{(\omega)} = \mathbb{E}_{b|y^*,\phi^{(\omega)}}(B'B)$, $R^{(\omega)} = \mathbb{E}_{b|y^*,\phi^{(\omega)}}(B'Z^*\tilde{D}b)$, and $T^{(\omega)} = \mathbb{E}_{b|y,\phi^{(\omega)}}(B)$. The matrix B is a stacked matrix of B_i, with each B_i a $T \times q(q-1)/2$ matrix, whose elements in each row are defined as $B_{ij} = (b_{jl} d_r z_{ijr} : l = 1, \ldots, (q-1), r = (l+1), \ldots, q)$, where B_{ij} denotes row j of the i^{th} matrix. P^- represents the Moore-Penrose inverse of P.

C Normal approximation test statistic

The test statistic B used to evaluate is derived as follows. The number of times N_f that the absolute forecast error of model one is smaller than a second model is a priori assumed to be binomial distributed with T_f repetitions and probability $p_{bin} = \frac{1}{2}$. Wackerly et al. (2014) state that a normal approximation for this distribution is valid if

$$T_f > 9 \cdot \max\left(\frac{p_{bin}}{1-p_{bin}}, \frac{1-p_{bin}}{p_{bin}}\right).$$

As the simulation study considers $T_f = 50$ and $\frac{p_{bin}}{1-p_{bin}} = \frac{1-p_{bin}}{p_{bin}} = 1$, the minimum condition is satisfied. This implies that I can assume N_f to be normally distributed with

$$\mathbb{E}(N_f) = p_{bin}T_f \qquad \text{, and} \qquad Var(N_f) = p_{bin}(1-p_{bin})T_f.$$

However, to make testing more convenient, I introduce the test statistic B. This statistic is the standardised version of N_f and thus defined as

$$\begin{aligned} B &= \frac{N_f - \mathbb{E}(N_f)}{\sqrt{Var(N_f)}} \\ &= \frac{N_f - p_{bin}T_f}{\sqrt{p_{bin}(1-p_{bin})T_f}}. \end{aligned}$$

Filling in the a priori probability $p_{bin} = \frac{1}{2}$ gives

$$\begin{aligned} B &= \frac{N_f - \frac{1}{2}T_f}{\sqrt{\frac{1}{4}T_f}} \\ &= \frac{2N_f - T_f}{\sqrt{T_f}} \end{aligned}$$

As the condition of N_f to be approximately normally distributed is $T_f > 9$ and B is N_f standardised and rewritten, I can take $T_f > 9$ as sufficient to assume $B \sim N(0,1)$.

D Tables

Table 6

Selecting Fixed and Random Effects

This table gives the number of times (in %) the covariates were correctly identified with each of the three variable selection methods in 100 simulated datasets, with a data generating process as in equation 20. Column one gives the variable that is changed compared to the baseline* scenario. Columns two to four give the percentage of cases where the model was correctly identified for variable y_1. Columns five to seven give the percentage of cases where the model was correctly identified for variable y_2.

Fixed Effects	y_1			y_2		
Scenario	Kokic	Chow	Bondell	Kokic	Chow	Bondell
Baseline*	16.6	59.8	39.0	11.2	61.0	69.0
$T = 50$	9.8	42.2	33.0	7.6	47.4	61.0
$T = 200$	20.4	59.4	0.0	7.2	67.4	0.0
$m = 10$	16.9	58.7	0.0	9.4	61.0	0.0
$m = 20$	18.1	59.3	0.0	8.5	62.6	0.0
$\sigma_\alpha = (3, 2, 3, 2)'$	16.0	60.8	8.0	5.6	59.2	78.0
$\beta = (5, 3, 2, 4)'$	14.2	61.8	2.0	9.0	63.4	28.0
$R_j = I_2$	22.2	56.4	56.0	9.4	56.2	63.0
$Q_j = I_4$	10.8	54.6	58.0	9.6	67.6	64.0
$u_{kt} \sim U(-2, 2)$	18.6	69.6	28.0	11.8	71.6	44.0
$u_{kt} \sim N(0, 1)$	19.6	73.0	44.0	7.8	73.8	39.0
Random Effects						
Baseline	13.8	57.8	6.0	11.0	54.8	50.0
$T = 50$	5.0	41.4	13.0	7.2	26.0	32.0
$T = 200$	20.0	59.4	0.0	7.2	59.8	0.0
$m = 10$	4.0	54.8	0.0	9.4	52.9	0.0
$m = 20$	15.0	57.6	0.0	8.4	54.6	0.0
$\sigma_\alpha = (3, 2, 3, 2)'$	12.4	58.6	3.0	5.4	52.2	76.0
$\beta = (5, 3, 2, 4)'$	10.0	60.0	0.0	9.0	56.4	23.0
$R_j = I_2$	17.6	46.8	2.0	9.2	40.4	26.0
$Q_j = I_4$	11.8	62.4	6.0	9.6	64.8	41.0
$u_{kt} \sim U(-2, 2)$	10.8	68.0	9.0	10.8	70.6	19.0
$u_{kt} \sim N(0, 1)$	9.8	70.4	16.0	7.8	72.8	16.0

* The baseline scenario has the following model parameters: $T = 100$, $m = 5$, $\sigma_\alpha = (1, 1, 1, 1)'$, $\beta = (1, 1, 1, 1)'$, $R_j = 0.4I_2$, $Q_j = 0.3I_2$. The covariates are generated from $u_{1t} \sim U(-1, 1)$, $u_{2t} \sim U(-0.5, 0.5) + 0.05(t - 0.5T)$, and $u_{3t} \sim U(-0.75, 0.75) + 1.25\cos(\pi t/6)$.

Table 7

Summary Statistics De Kooy

This table shows the summary statistics for the dependent and explanatory variables used in this study. The values are reported for the weather station in De Kooy in the sample period January 1959 through December 2017. Rows one to four show the pre-transformed dependent variables and rows five to nine show the five explanatory variables.

Variable	Mean	Std. Dev.	Min	Max
$\text{logit}(p_{rain})$	−0.677	0.800	−4.595	1.872
$\log(\mu_{rain})$	5.457	2.134	0.000	14.862
$\Delta_{12}Temp_{max}$	0.034	2.249	−8.100	7.000
$\Delta_{12}Temp_{min}$	0.028	2.168	−9.400	8.400
CO_2 level	353.481	26.713	313.260	409.650
NAO index	0.013	1.020	−3.180	3.040
ρ_{cyc}	3.284	1.219	0.000	7.000
$\rho_{anti\text{-}cyc}$	3.439	1.222	0.000	7.000
RH	0.833	0.043	0.714	0.948

Table 8

Summary Statistics Eelde

This table shows the summary statistics for the dependent and explanatory variables used in this study. The values are reported for the weather station in Eelde in the sample period January 1959 through December 2017. Rows one to four show the pre-transformed dependent variables and rows five to nine show the five explanatory variables.

Variable	Mean	Std. Dev.	Min	Max
$\text{logit}(p_{rain})$	−0.588	0.765	−4.595	1.427
$\log(\mu_{rain})$	5.491	1.951	0.000	15.800
$\Delta_{12}Temp_{max}$	0.031	2.623	−8.600	8.600
$\Delta_{12}Temp_{min}$	0.011	2.389	−10.400	10.100
CO_2 level	353.481	26.713	313.260	409.650
NAO index	0.013	1.020	−3.180	3.040
ρ_{cyc}	3.295	1.204	0.000	7.000
$\rho_{anti\text{-}cyc}$	3.448	1.196	1.000	7.000
RH	0.849	0.057	0.683	0.959

Table 9

Summary Statistics Vlissingen

This table shows the summary statistics for the dependent and explanatory variables used in this study. The values are reported for the weather station in Vlissingen in the sample period January 1959 through December 2017. Rows one to four show the pre-transformed dependent variables and rows five to nine show the five explanatory variables.

Variable	Mean	Std. Dev.	Min	Max
logit(p_{rain})	−0.707	0.746	−4.595	1.649
log(μ_{rain})	5.487	2.116	0.000	17.814
$\Delta_{12}Temp_{max}$	0.030	2.272	−7.800	7.500
$\Delta_{12}Temp_{min}$	0.026	1.958	−8.600	8.100
CO_2 level	353.481	26.713	313.260	409.650
NAO index	0.013	1.020	−3.180	3.040
ρ_{cyc}	3.274	1.286	0.000	7.000
$\rho_{anti\text{-}cyc}$	3.475	1.261	1.000	7.000
RH	0.818	0.049	0.668	0.939

Table 10

Summary Statistics Beek

This table shows the summary statistics for the dependent and explanatory variables used in this study. The values are reported for the weather station in Beek in the sample period January 1959 through December 2017. Rows one to four show the pre-transformed dependent variables and rows five to nine show the five explanatory variables.

Variable	Mean	Std. Dev.	Min	Max
logit(p_{rain})	−0.692	0.718	−4.595	1.056
log(μ_{rain})	5.660	2.115	0.000	17.873
$\Delta_{12}Temp_{max}$	0.034	2.778	−8.800	10.000
$\Delta_{12}Temp_{min}$	0.013	2.304	−9.600	9.300
CO_2 level	353.481	26.713	313.260	409.650
NAO index	0.013	1.020	−3.180	3.040
ρ_{cyc}	3.274	1.314	0.000	7.000
$\rho_{anti\text{-}cyc}$	3.544	1.260	1.000	8.000
RH	0.804	0.068	0.598	0.948