Local temperature forecasts based on statistical post-processing of numerical weather prediction data

Alerskans, Emy; Kaas, Eigil

Published in:
Meteorological Applications

DOI:
10.1002/met.2006

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Local temperature forecasts based on statistical post-processing of numerical weather prediction data

Emy Alerskans¹,² | Eigil Kaas¹,³

¹The Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
²FieldSense A/S, Aarhus, Denmark
³Danish Meteorological Institute, Copenhagen, Denmark

Abstract
Six adaptive, short-term post-processing methods for correcting systematic errors in numerical weather prediction (NWP) forecasts of near-surface air temperatures using local meteorological observations are assessed and compared. The methods tested are based on the simple moving average and the more advanced Kalman filter. Forecasts from the rather coarse-resolution global NWP model Global Forecast System (GFS) and the regional high-resolution NWP model HARMONIE are post-processed, and the results are evaluated for 100 private weather stations in Denmark. The performance of the post-processing methods differs depending on the NWP model. Overall, the combined moving average and a so-called lead time Kalman filter performs the best. The moving average was shown to be superior to a diurnal bias correction Kalman filter at removing the longer-term systematic errors for HARMONIE forecast data and comparable for GFS forecast data. Subsequent application of the lead time Kalman filter corrects for the short-term errors using the real-time forecast error. The post-processing method is adaptive and there is no need for a long record of observations or a historical archive of forecasts.

KEYWORDS
Kalman filter, local weather forecasts, NWP, post-processing

1 | INTRODUCTION
Weather forecasts are important for, for example, agricultural systems since the weather has a significant impact on the growth, development and yield of crops as well as the water and fertilization need (Petr, 1991; Stigter et al., 2000; WMO, 2010). Hence, accurate weather forecasts can have an impact on activities related to agricultural systems (Hoogenboom, 2000; Pelosi et al., 2016). The present study has been motivated by the need for more accurate and localized forecasts in relation to agricultural activities.

It is a well-known fact that numerical weather prediction (NWP) models exhibit systematic errors, especially for near-surface variables such as air temperature and wind speed. This is partly due to limitations to initial conditions, deficiencies in the physical formulation of the model dynamics and the inability of these models to successfully handle sub-grid phenomena (Krishnamurti et al., 2004; Mass et al., 2002; Nicolis et al., 2009; Paegle et al., 1997;
Tribbia & Baumhefner, 2004). In addition, the output from NWP models is gridded, which means that it represents an average over a grid point/cell (Pielke, 2013). Even though high-resolution regional NWP models usually have a spatial resolution of a few kilometres (or even finer), they generally exhibit local biases due to unresolved topography and obstacles (Hart et al., 2005).

Statistical post-processing methods are commonly used to correct systematic forecast errors. The perfect prog (short for perfect prognosis) (Klein et al., 1959) and model output statistics (MOS) (Glahn & Lowry, 1972) are two of the first and most commonly used statistical post-processing methods in meteorology. Perfect prog relates forecasts to observations using linear regression, assuming that the forecast is perfect. This means that it cannot correct for possible dynamical model errors or biases (Wilks, 2011). MOS is usually preferred over perfect prog since it uses the forecasts as predictors also in the development of the regression equations, which means that it can include the model errors and biases in the regression (Wilks, 2011). The major disadvantage of MOS is that it requires a long training dataset, at least 2 years of historical NWP forecasts, to develop useful MOS equations (Jacks et al., 1990). Furthermore, it also requires the NWP forecasting system not to undergo any changes during the training and prediction period (Kalnay, 2003).

Adaptive post-processing methods with a shorter training period have also been used to correct systematic errors in NWP forecasts. Previous studies have shown that post-processing by moving averages (MAs) can improve upon the raw NWP forecast (Eckel & Mass, 2005; McCollor & Stull, 2008; Stensrud & Skindlov, 1996; Stensrud & Yussouf, 2005; Woodcock & Engel, 2005). Another adaptive post-processing method with a short training period that has shown promising results is the Kalman filter (Kalman, 1960). Among the first to use the Kalman filter for post-processing of meteorological variables were Persson (1991) and Simonsen (1991), who used the Kalman filter to post-process temperature forecasts. Since then, the Kalman filter has been applied to correct the diurnal bias in the forecast of 2 m temperatures (Homleid, 1995), to adjust maximum and minimum temperature forecasts (Galanis & Anadranistakis, 2002) and to post-process near-surface temperature, humidity and wind speed forecasts (Galanis et al., 2006; Libonati et al., 2008; Louka et al., 2008; Sweeney et al., 2013).

The objective of this study is to compare different post-processing methods and how well they can downscale state-of-the-art NWP forecasts to obtain localized forecasts. The focus is on how well the different post-processing methods perform with respect to post-processing of a high-resolution NWP model and a coarser-resolution NWP model for a region with relatively homogeneous topography. Six different post-processing methods are applied to 2-day NWP forecasts of 2 m temperature from the Global Forecast System (GFS) model, which is a global NWP model with a relatively coarse spatial horizontal resolution, and the HARMONIE model, which is a regional NWP model with a high spatial horizontal resolution. The post-processing methods are based on the MA and the Kalman filter, aiming at eliminating recent operational model biases. The results are evaluated for 100 private weather stations (PWSs) in Denmark for the period 1 October 2019 to 31 May 2020. Since the network of weather stations is relatively new, no long record of observations is available. Hence, we have focused on post-processing methods that do not require a long training period.

Previous studies have also compared different post-processing methods. Cheng and Steenburgh (2007) compared the post-processing of 2-m temperature, 2-m dew point and 10-m wind speed forecasts from an intermediate-resolution model using MOS, an MA and a Kalman filter. Anadranistakis et al. (2004) compared post-processing by a Kalman filter with that of an empirical post-processing method on 2-m temperature and humidity forecasts from a fine-resolution model, with focus on an agricultural site in Greece. Vashani et al. (2010) compared several methods for post-processing maximum and minimum surface temperatures, also focusing on agricultural sites, but used an intermediate-resolution model. As stated above, we have chosen to compare different variants of the Kalman filter and the MA since these do not require a long training period. We have chosen to compare post-processing methods that focus on reducing both the longer-term, diurnal systematic errors and the very short-term systematic errors. Furthermore, we have focused on comparing the results from post-processing on both a coarse-resolution and a high-resolution NWP model in order to analyse any difference in performance based on model resolution. In addition, our focus is on agricultural sites in Denmark. Compared to the areas of interest for the above-mentioned studies, the terrain in Denmark is much less complex; except for land–sea contrasts it is generally flat and homogeneous. Therefore, the type of systematic errors is likely to be coupled to different aspects of the models.

The paper is structured such that Section 2 describes the observational and forecast data used. Section 3 presents the post-processing methods, and the results are shown in Section 4. The results and future prospects are discussed in Sections 5, and Section 6 summarizes the paper with conclusions.

2.1 Observational data

Crowd-sourced data are commonly defined as data obtained through outsourcing to citizens (Howe, 2006).
Due to recent technological advances, the term “crowd-sourced data” also includes data obtained from private sensors (Muller et al., 2015). In atmospheric sciences, an increasing number of data obtained through crowd-sourcing from private sensors are being investigated, such as data from smartphones (Hintz et al., 2019; Kim et al., 2016) and PWSs (Chapman et al., 2017; Nipen et al., 2020). As opposed to official stations, PWSs usually do not comply with the rules and standards defined by the World Meteorological Organization (WMO) (WMO, 2018). This means that PWSs can be placed in areas that are not well-suited for measurements of meteorological variables, such as close to buildings and trees, on balconies or even indoors. Another disadvantage of crowd-sourced data from private sensors is that the measurement instruments usually are low-cost and not as accurate and well-calibrated as the instruments used in official weather station networks (Bell et al., 2013, 2015). The major advantage of crowd-sourced data, on the other hand, is the spatial density of observations, where observations from non-traditional sources vastly outnumber observations from official networks.

The observational data used in this study consist of observations from FieldSense’s network of PWSs. Other providers of PWSs also exist, such as Netatmo. FieldSense has approximately 1000 stations in Denmark as of 20 October 2020, whereas the national meteorological service in Denmark, the Danish Meteorological Institute (DMI), has approximately 70 WMO-compliant stations (although not all of them measure temperature). This means that the private network owned by FieldSense is more than 10 times as dense as the network operated by DMI. The weather stations from FieldSense are mainly developed for use in agricultural applications and are generally placed on or in the vicinity of fields. Hence, the network coverage is densest in areas with a high agricultural activity. Netatmo has an even denser network of PWSs compared with FieldSense. However, their coverage is densest in populated areas, such as cities, and more sparse in remote and rural areas. This is the main reason for choosing FieldSense’s network in the first place, since the aim of this study is to develop a post-processing method for correcting local systematic errors in an NWP forecast for the benefit of agricultural activities. Furthermore, since the aim is to produce site-specific forecasts that better match the locally observed weather, it is assumed that the observations represent the truth, that is, as a fundamental premise for the entire study we are not aiming at forecasting the true local temperature but “only” the actual temperature recorded. Therefore, the forecasts are validated using the observations and comparisons between the post-processed and raw forecasts are based on their performance relative to the observations.

The observational data consist of 2 m air temperature observations from a selection of 100 FieldSense stations in Denmark. Figure 1 shows the location of the selected stations. The temporal resolution of the observations is 10 min; however, here we have only used the observations closest in time to the output times of the NWP models used. It is important to apply appropriate quality control (QC) methods to identify obviously erroneous observations. In this study, we have focused on using QC methods based on the temporal properties of the individual stations. These tests consist of a plausibility check, a rate of change check and a persistence check (Fiebrich et al., 2010; Zahumenský, 2004). Furthermore, since the stations can be moved, only stations that were stationary throughout the whole investigation period were used. In addition, indoor locations were excluded.

2.2 | NWP data

As previously mentioned, the post-processing methods were tested on forecast data from two different NWP models, namely GFS from the National Center for Environmental Prediction (NCEP) and the DMI-HARMONIE NWP model, which is developed by the HIRLAM and ALADIN consortia and run operationally by the Icelandic and Danish Meteorological Institutes (IMO and DMI). The two models were chosen to facilitate a comparison between post-processing of forecasts from a coarse-resolution model and a high-resolution model.

GFS is a hydrostatic global NWP model with a finite-volume dynamical core with a horizontal resolution of approximately 13 km and 64 vertical layers. The forecast is updated four times a day (00, 06, 12 and 18 UTC) and

![Figure 1](image_url) Location of the 100 private weather stations used in this study. The stations marked by a blue star indicate the locations of the stations at Ringsted (1), Langeland (2), Faaborg (3) and Tørring (4)
has hourly forecast output for the first 120 h, whereas forecast output for every 3 h is produced for forecast days 5–16. In this study, the GFS data with a spatial resolution of 0.25° were used (NCEP, 2015). The data were accessed via the National Center for Atmospheric Research (NCAR) Research Data Archive (https://rda.ucar.edu/). Since historical GFS data were used, which only include model output for every third hour, the temporal resolution of the GFS data used in this study is three-hourly.

HARMONIE is a non-hydrostatic limited-area NWP model operated at 2.5 km horizontal resolution with 65 vertical layers. A detailed description of the HARMONIE model configuration can be found in Bengtsson et al. (2017), and Yang et al. (2017) the operational implementation of the current version of HARMONIE at DMI. The forecast is updated eight times a day and has lead times up to 60 h with hourly output.

In this study, forecast data for up to 2 days ahead were extracted for both models. Data were extracted for the period 1 October 2019 to 31 May 2020 for Denmark. The forecast data from both models were interpolated to the station locations using bi-linear horizontal interpolation of the four closest grid points. The performance of the post-processing methods and the raw NWP forecast were tested on a subset of the data using both bi-linear horizontal interpolation and the nearest grid point in order to investigate the effect of interpolation to station location. Generally, using the nearest grid point resulted in slightly higher errors. The improvement obtained from the post-processing methods, however, were very similar, regardless of the interpolation method used.

3 | STATISTICAL POST-PROCESSING METHODS

Six post-processing methods were evaluated and compared in this study. The methods considered range from the simple MA to the more advanced Kalman filter. They are all adaptive, meaning that they correct the current forecast using previous forecasts and local observations.

3.1 | Moving average

An MA is tested. Separate MAs are used for each forecast cycle and they are constructed to remove the bias based on forecast lead time. The mean error for each forecast lead time over a specified window is calculated and the results are then subtracted from the current forecast. The window over which to calculate the MA was investigated for a subset of the data and results for window sizes in the range 3–30 days were compared for both forecast datasets. For the GFS forecast data, the root mean square error (RMSE) of the post-processed forecast was reduced with increasing window length. The differences between the window lengths were most pronounced for the longer lead times, where the 30-day MA clearly performed the best. For HARMONIE forecast data, using a window length of less than 10 days resulted in an increase in RMSE for the longer lead times. Again, the 30-day MA gave the best results, especially for the longer lead times. Therefore, the window size was set to 30 days.

3.2 | Weighted moving average

The weighted moving average (WMA) puts higher weights on the most recent observations, as opposed to the simple MA, which weights all observations equally. A linearly WMA, also constructed to correct for lead time biases, was tested. The window size was set to the same as that of the simple MA, namely 30 days.

3.3 | Kalman filter

The Kalman filter is an optimal recursive algorithm for estimating an unknown process based on observations and information about the process’ temporal evolution. In this section, the general form of the Kalman filter equations is described. For a more detailed and complete description see, for example, Kalman (1960) and Gelb (1974).

The system equation and the observation equation form the basis of the Kalman filter. The system equation describes the relationship between the state vector of the unknown process, \( \mathbf{x}_t \), at time \( (t-1) \) and at time \( t \), that is, it describes the prediction of \( \mathbf{x}_t \) from time \( (t-1) \) to time \( t \)

\[
\mathbf{x}_t = \mathbf{F}_t \mathbf{x}_{t-1} + \mathbf{B}_t \mathbf{u}_t + \mathbf{G}_t \mathbf{w}_t.
\]

Here \( \mathbf{x}_t \) is an \( n \times 1 \) vector, describing the state of the process at time \( t \), where \( n \) is the dimension of the filter, which varies depending on how the Kalman filter is implemented. In the present study, the state vector is the true forecast error of the 2 m temperature \( (T_{2m}) \), defined as forecast minus observed temperature, and \( n \) is given either by the forecast length or the number of forecast hours per day, depending on the filter implementation (see Sections 3.3.1 and 3.3.2). \( \mathbf{w}_t \) is a Gaussian, zero-mean white noise process with variance \( Q_t \); \( \mathbf{w}_t \sim N(0,Q_t) \), describing the random change in the evolution of the forecast error from time \( (t-1) \) to \( t \) and \( \mathbf{u}_t \) is a control input, which represents any input added to the system. Both of these vectors also have dimensions \( n \times 1 \). \( \mathbf{F}_t \) is the
system matrix, which describes how the \( T_{2m} \) forecast error evolves with time, \( G_t \) is the system noise matrix, which is related to the uncertainty of the system, and \( B_t \) is the input matrix, which is related to the control input. These matrices all have dimensions \( n \times n \).

The observation equation relates the \( T_{2m} \) forecast error (i.e., the state vector \( x_t \)) to the observed 2 m temperature forecast error, \( y_t \),

\[
y_t = H_t x_t + v_t.
\]  
(2)

Here, \( y_t \) is the observation vector, which contains the observations and has dimensions \( p \times 1 \), where \( p \) is the number of observations at each given time. In this study, \( p = 1 \) since we are only considering one observation per station for each given time. \( H_t \) is the observation matrix, with dimensions \( p \times n \), which maps the state vector to the observation space, and \( v_t \) is a Gaussian, zero-mean white noise process with variance \( R_t \); \( v_t \sim N(0,R_t) \), describing the random observation error.

The Kalman filter is an iterative method and can be divided into two steps: the prediction step and the update step. In the first step, the prediction equations are used to predict the estimates of the state vector, \( \hat{x}_{t|t-1} \), and its error covariance, \( \hat{P}_{t|t-1} \), ahead in time, based only on the previous estimates of the state, \( \hat{x}_{t-1|t-1} \), and its error covariance, \( \hat{P}_{t-1|t-1} \),

\[
\begin{align*}
\hat{x}_{t|t-1} &= F_t \hat{x}_{t-1|t-1} + B_t u_t, \\
\hat{P}_{t|t-1} &= F_t \hat{P}_{t-1|t-1} F_t^T + G_t Q_t G_t^T.
\end{align*}
\]  
(3)
(4)

After the prediction step follows the update step, where observations from time \( t \) are used to update the estimate of the state vector and its error covariance

\[
\begin{align*}
\hat{x}_{t|t} &= \hat{x}_{t|t-1} + K_t (y_t - H_t \hat{x}_{t|t-1}), \\
\hat{P}_{t|t} &= (I - K_t H_t) \hat{P}_{t|t-1},
\end{align*}
\]  
(5)
(6)

where \( K_t \) is the Kalman gain, given by

\[
K_t = \hat{P}_{t|t-1} H_t^T (H_t \hat{P}_{t|t-1} H_t^T + R_t)^{-1}.
\]  
(7)

The difference \( y_t - H_t \hat{x}_{t|t-1} \) in Equation (5) is called the innovation (or residual). The Kalman gain determines how much weight is put on the previous estimate and how much is put on the innovation. In other words, the Kalman gain determines how fast the filter can adapt to new observed changes.

Equations (3)–(7) are used to update the Kalman filter from time \( t-1 \) to \( t \). Prior to running the filter, the following must be defined: the system matrix \( F_t \), the observation matrix \( H_t \), the input matrix \( B_t \), the noise matrix \( G_t \), the system noise covariance matrix \( Q_t \) and the observation noise covariance matrix \( R_t \). Furthermore, initial values for the state vector, \( x_0 \), and the state covariance matrix, \( P_0 \), must be given. The choice of \( x_0 \) and \( P_0 \) is not very important for the performance of the filter since their estimates will converge relatively fast towards their “true” Kalman-estimated values (Welch & Bishop, 2002). However, the choice of noise covariances affects the outcome of the filter significantly. Since only one temperature observation per station is available at any time, the covariance of the observation noise, \( R_t \), is a scalar, which can be written as \( R_t = R^2 \), where \( R^2 \) is the variance of the observation noise. The covariance matrix of the system noise, on the other hand, can be written as the product between the variance \( Q^2 \) and a correlation matrix, \( r_{corr} \); \( Q = Q^2 r_{corr} \). The ratio \( Q/R \) affects the Kalman gain, and therefore also how fast the filter is able to adapt to new changes. How to choose the covariance matrices of the system and observation noises is one of the most difficult aspects of constructing a Kalman filter. Several different methods have been proposed, such as tuning to obtain a desired performance or property (Homleid, 1995) and online estimations (Crochet, 2004; Galanis & Anadranistakis, 2002). The method adopted here is an adaptive method where the structure of the system noise matrix, \( r_{corr} \), is updated continuously based on previous forecast errors, whereas the ratio \( Q/R \) is based on tuning to obtain a good performance of the filter. The calculation of \( r_{corr} \) is described in Section 3.3.3. The time step update, that is, how often the Kalman filter is run, is the same as the output frequency of the NWP model, that is, 1 h for HARMONIE forecasts and 3 h for GFS forecasts.

Two different variants of the Kalman filter are tested for post-processing of the 2 m temperature forecasts. They are described more in detail in the following.

### 3.3.1 Diurnal bias correction Kalman filter

The first variant of the Kalman filter tested here is a diurnal bias correction Kalman filter (KFDBC) that is based on the Kalman filter proposed by Homleid (1995). The aim of this filter is to correct systematic diurnal forecast errors and therefore the state vector in the Kalman filter is a vector containing one element for each forecast output hour of the day. Since the forecast errors are assumed to vary only during a 24-h period, forecasts valid at the same time will be corrected equally. As an example, a forecast
valid at +01 h and +25 h will be corrected by the same correction term.

To fully describe the KF_{DBC} system, the following matrices in Equations (3)–(7) need to be defined: $F_t$, $H_t$, $Q_t$, $R_t$, $B_t$ and $G_t$, as well as the initial conditions $x_0$ and $P_0$.

Since there is no simple relationship describing how $x_t$ evolves with time, the system matrix is defined as $F_t = I_n$. $I_n$ is the identity matrix and $n$ denotes the dimension of the filter ($n = 24$ for HARMONIE and $n = 8$ for GFS). The observation matrix, $H_t$, is a $1 \times n$ vector. Since the filter should make corrections on a diurnal basis, the observation matrix varies with time, such that only the observation from the current time $t$ is considered by the filter. For example, at 00 UTC the observation matrix for application to HARMONIE forecasts is $H_t = [1 \ 0 \ 0 \ \cdots \ 0]$, and at 01 UTC it is $H_t = [0 \ 1 \ 0 \ \cdots \ 0]$, and so forth.

Different values for $Q$ and $R$ were tested and it was found that the optimal value for the ratio $Q/R$ was 0.03. As mentioned, this ratio decides how fast the Kalman filter adapts to new changes. The larger the ratio $Q/R$, the faster the Kalman filter adapts to new changes. However, when investigating the performance of the Kalman filter for different ratios, it was found that higher ratios lead to higher variability in the corrections of KF_{DBC} and therefore also a larger standard deviation of the post-processed forecast, which is not desirable. This was also shown by Homleid (1995). Therefore, the choice was made to use a ratio of 0.03.

No control input is used and therefore $B_t$ is not needed. In addition, the system noise matrix, $G_t$, is taken as the identity matrix since the system noise is considered time-invariant. The initial value for $x_t$ is chosen as $x_0 = [0 \ 0 \ \cdots \ 0]^T$, where $x_t$ is an $n \times 1$ vector. Since the estimate of $x_t$ converges relatively fast, it is of little importance how the state is initialized (Welch & Bishop, 2002). For the initial value of $P_t$, the structure is chosen identical to the correlation matrix of the system noise covariance. The variance of $P_0$ is chosen as $P^2 = 4$, to indicate the uncertainty associated with the initial guess for $x_t$.

### 3.3.2 | Lead time Kalman filter

The lead time Kalman filter (KF_{LT}) is constructed for forecast lead time instead of time of the day and is based on the local adaptive Kalman filter proposed by Doeswijk and Keesman (2005). A brief description of the relevant parameters is given here; for further details on the filter, see the study by Doeswijk and Keesman (2005).

The state vector, $x_t$, represents the $T_{2m}$ forecast error and has dimensions $n \times 1$, where $n$ is the dimension of the filter and is equal to the number of output per forecast, that is, $n = 48$ for HARMONIE and $n = 16$ for GFS. For update times where no new forecast is available, the state vector is updated so that the first element always represents the next lead time in the forecast. However, whenever a new forecast is available, at time $t'$, the state vector is reset since it is assumed that the new forecast is better than the KF-updated old forecast. Therefore, $F_t$ is an $n \times n$ time-varying matrix given by

$$F_t = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \quad \forall t \neq t'$$

(8)

$$F_t = 0 \quad \forall t = t'$$

(9)

The observation matrix is time-invariant and defined so that only the most recent observation is considered; $H_t = [1 \ 0 \ 0 \ \cdots \ 0]$. Different values for $Q$ and $R$ were tested. The desired properties for the lead time Kalman filter is that it should be able to quickly adjust the forecast in the very near future, which means that the ratio $Q/R$ should be relatively large. By testing different values for $Q$ and $R$, it was found that the optimal ratio is $Q/R = 6.0$.

$G_t$ is an $n \times n$ time-varying matrix, which is only relevant at time $t'$, that is, whenever a new forecast is available. The same applies to the input matrix $B_t$. Therefore

$$G_t = B_t = I_n \quad \forall t = t'$$

(10)

$$G_t = B_t = 0 \quad \forall t \neq t'.$$

(11)

The initial value for $x_t$ is chosen as $x_0 = [0 \ 0 \ \cdots \ 0]^T$ and the initial value of the structure of $P_0$ is chosen identical to the correlation matrix of the system noise covariance. To reflect the uncertainty associated with the initial guess for $x_t$, the variance of $P_0$ is chosen as $P^2 = 4$.

### 3.3.3 | System noise covariance matrix

The system noise covariance matrix can be decomposed into two parts: a variance, $Q^2$, and a correlation matrix, $r_{corr}$. It is, as mentioned, the ratio between $Q$ and $R$ that decides how fast the Kalman filter will react to new changes in the bias. On the other hand, it is the structure of $r_{corr}$ that decides how the Kalman correction is propagated from the current update time to the rest of the
Rearrangements between forecast hours, \( r_{corr} \) is updated continuously. The previous \( N \) forecast–observation pairs are used to calculate the auto-correlation of the \( T_{2m} \) forecast error

\[
r_k = \frac{\sum_{i=1}^{N-k} (x_i - \bar{x})(x_{i+k} - \bar{x}_{lagged})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2 \sum_{i=k+1}^{N} (x_i - \bar{x}_{lagged})^2}},
\]

where \( x_i \) is the ith forecast error for the original data and \( x_{i+k} \) is the ith forecast error for the \( k \)-unit lagged data.

The number of forecast–observation pairs, \( N \), to use in the calculation was investigated, where 5, 10, 15, 20, 25 and 30 days (where each day consists of data from all forecast cycles) were tested. Using only 5 days of data resulted in a too variable correlation matrix with time, which lead to an increase in the standard deviation of the post-processed forecast over time. A window size of 30 days resulted in a more smooth yet still varying correlation matrix, which did not result in an increase in the standard deviation of the post-processed forecast. As a result, 30 days of forecast–observation pairs are used to calculate the correlation matrix for all Kalman filter post-processing algorithms. The correlation matrix is updated once every day.

The \( Q/R \) ratios used for KFDBC and KFLT, defined in Sections 3.3.1 and 3.3.2, were optimized based on the choice of a window size of 30 days for the calculation of the correlation matrix.

3.4 Combined post-processing methods

3.4.1 Diurnal bias correction Kalman filter followed by lead time Kalman filter

NWP bias usually has both a diurnal component and a real-time component. The diurnal bias correction Kalman filter aims to correct the systematic diurnal forecast errors, whereas the lead time Kalman filter aims at correcting the forecast error in the very short-range time period. By applying the lead time Kalman filter after the diurnal bias correction Kalman filter, a combination between diurnal corrections and very short-range corrections would be obtained. Therefore, the fifth post-processing algorithm, referred to as the double Kalman filter (KF2x), is a combination of the diurnal bias correction Kalman filter followed by the lead time Kalman filter.

3.4.2 MA followed by lead time Kalman filter

The MA is also constructed to remove the longer-term bias. Hence, a logical combination after having combined the diurnal bias correction Kalman filter and the lead time Kalman filter is to combine the lead time Kalman filter with the MA as well. The last post-processing method, MA + KFLT, is therefore a superposition of the MA and the lead time Kalman filter.

4 RESULTS

The six post-processing methods described in Section 3 are applied to both GFS and HARMONIE \( T_{2m} \) forecast data for the period 1 October 2019 to 31 May 2020. The Kalman filter tends to converge relatively fast towards its “true” Kalman estimates (Welch & Bishop, 2002). The MA-type post-processing methods, however, have a window size of 30 days. Therefore, the month of October is considered as training period for all post-processing methods and is discarded in the statistical analysis. To assess the performance of the post-processing methods, the following verification scores are used

\[
\text{Mean error (ME)} / \text{Bias} = \frac{1}{N} \sum_{i=1}^{N} (y_i - o_i),
\]

\[
\text{Mean absolute error} : \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - o_i|,
\]

\[
\text{Standard deviation} : \text{STD} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - o_i - \text{ME})^2},
\]

\[
\text{Root mean square error} : \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - o_i)^2},
\]

where \( N \) is the number of observation–forecast data pairs, \( y \) denotes NWP forecast data and \( o \) denotes observational data, which are assumed to represent the ground truth. It should be noted that we do not consider any eventual instrument bias at all in the very local observations, and validate the results against the actual recorded 2 m temperatures. Furthermore, skill scores based on the above-mentioned verification metrics are used to evaluate
the performance of the post-processed forecast relative to the performance of the raw NWP model output

\[
SS_X = 1 - \frac{X_{PP}}{X_{NWP}},
\]

where \(X_{PP}\) refers to verification metric \(X\) for post-processing method \(PP\), \(X_{NWP}\) refers to the corresponding verification metric for the raw NWP model output and \(SS_X\) refers to the skill score for the same verification metric.

The post-processing methods were applied to observational data from 100 weather stations; however, detailed statistics are only showed for four selected stations, namely Ringsted, Langeland, Faaborg and Tørring stations. The location of these four stations is marked with a blue star in Figure 1. The statistics have been calculated for the whole time period (excluding October), using all forecast cycles and all updated forecasts.

Figures 2–4 show the bias, mean absolute error (MAE) and standard deviation (STD) of the post-processed GFS forecast for the four example stations. The raw GFS forecast shows a positive bias for all four stations, with Tørring station having the largest overall bias, 0.85°C. It should be noted that the bias of the raw GFS data decreases slightly with lead time for all stations, which indicates a small cold drift for the GFS model. Tørring station also exhibits the largest overall MAE, 1.21°C. However, the largest overall STD of the raw GFS forecast is seen for Langeland station, 1.70°C. Both Langeland and Faaborg stations exhibit small biases, whereas the MAE shows larger absolute errors. For the post-processing methods, application of KF\(_{LT}\) results in a decrease in bias for mostly the shorter lead times for all four stations. The lead time Kalman filter also reduces both the MAE and STD mostly for the shorter lead times. KF\(_{DBC}\) and KF\(_{X2}\) reduce the bias close to zero for almost all lead times. However, the biases of these two post-processing methods decrease slightly with lead time, similar to the behaviour of the raw GFS model bias. The MAE and STD are reduced for all stations and lead times when using KF\(_{DBC}\) and KF\(_{X2}\). The other three post-processing methods, MA + KF\(_{LT}\), MA and WMA, reduce the bias of the forecast close to zero for most lead times. They also reduce the MAE and STD for all stations and all lead times. KF\(_{X2}\) and MA + KF\(_{LT}\) perform best for the shorter lead times, whereas the performance of KF\(_{DBC}\), MA and WMA are comparable to the two combined post-processing methods for the longer lead times.

The corresponding results for post-processing of HARMONIE forecast data are shown in Figures 5–7. The raw HARMONIE forecast has small biases for Ringsted, Langeland and Faaborg stations, whereas the bias for Tørring station is higher: 0.50°C on average. The raw HARMONIE forecast exhibits a more pronounced decrease in bias with forecast lead time for all stations:

![Figure 2](image-url)
up to almost 0.5°C over the whole forecast length for Tørring station. The overall MAE and STD for all stations are similar: 0.76–0.99°C and 1.05–1.43°C, respectively. As was observed for the GFS forecast data, application of KF<sub>LT</sub> mostly affects the shorter lead times also when applied to the HARMONIE forecast data: up to +9 h. Using KF<sub>DBC</sub> and KF<sub>X2</sub> for post-processing results in a decrease in the bias with lead time for all stations, with a shape very similar to that of the raw HARMONIE model bias. On the other hand, the MAE is increased for
all four stations for the longer lead times when applying these two post-processing methods. KFX2 does, however, decrease the MAE for the very short lead times. The standard deviation is also decreased, mostly for the shorter lead times. MA + KF_{LT}, MA and WMA are the only post-processing methods that do not show the same decrease in bias as the raw NWP model bias with lead time. Instead, they reduce the bias closer to zero for all lead times. The two moving-average type post-processing methods perform similar to KF_{DBC} for the shorter lead times, and better for the longer lead times. On the other hand, they do not reduce the STD as much for the shorter lead times. Overall, MA + KF_{LT} is the post-processing method that reduces the MAE and STD the most for the shorter lead times, whereas it performs comparable to the other post-processing methods for the longer lead times.

A zigzag type of pattern can be seen for some of the stations, for both GFS and HARMONIE forecast data (see, e.g., Figures 3 and 5). The reason for this oscillating pattern is a difference in diurnal amplitudes between the forecasts and observations. When the statistics are calculated and presented as a function of forecast lead time, the difference in diurnal amplitude between forecasts and observations for the different forecast cycles will give rise to the type of zigzag pattern shown here.

The summary statistics—average bias, in absolute values, and average STD—based on all 100 stations are shown in Figures 8 and 9. For the post-processing methods, it can be seen that it is a general feature that KF_{LT} reduces the bias mostly for the shorter lead times, for both GFS and HARMONIE forecast data. However, the reduction in bias is generally seen for longer lead times for the GFS forecasts. For KF_{DBC} and KFX2, a slight increase in bias can be seen for the GFS forecast data. The increase is much more evident for HARMONIE forecast data, where these two post-processing methods in general tend to increase the bias for the longer lead times so that it is higher than that of the raw HARMONIE forecasts. The two moving-average type post-processing methods and MA + KF_{LT} generally produce forecasts with biases closer to zero for all lead times. As seen before, the effect of applying KF_{LT} is seen mostly for the shorter lead times, also for STD. The other post-processing methods generally reduce the STD for all lead times for GFS forecast data. For HARMONIE forecast data, however, only the lead time Kalman filter and the two combined post-processing methods reduce the STD and only for the short lead times.

It is often the larger errors that have an impact on the end-users and therefore, the post-processing methods are also evaluated based on how well they reduce the fraction of forecast busts, defined as forecast errors greater than 3°C. Figure 10 shows the fraction of forecast busts as a function of lead time for all stations. The fraction of forecast busts increases with lead time for the raw GFS
forecast and is in the range 5%–7%. KFLT reduces the fraction of forecast bust the least and mostly for the shorter lead times. The other post-processing methods reduce the fraction of forecast busts for all lead times by 3–4 percentage points. The fraction of forecast busts for the raw HARMONIE forecast also increases with lead time and is in the range 1.5%–3.5%. A smaller reduction in forecast busts after post-processing is evident for the
FIGURE 8  Bias summary statistics for all 100 stations as a function of lead time. Results are shown for post-processing of (a) Global Forecast System (GFS) and (b) HARMONIE forecasts for the diurnal bias correction Kalman filter (KF_DBC; blue), the lead time Kalman filter (KF_LT; light green), the double Kalman filter (KF_X2; magenta), the combined moving average and lead time Kalman filter post-processing method (MA + KF_LT; red), the 30-day moving average (MA; green), the 30-day linearly weighted moving average (WMA; orange) and the raw numerical weather prediction (NWP) model output (NWP; black).

FIGURE 9  Standard deviation summary statistics for all 100 stations as a function of lead time. Results are shown for post-processing of (a) Global Forecast System (GFS) and (b) HARMONIE forecasts for the diurnal bias correction Kalman filter (KF_DBC; blue), the lead time Kalman filter (KF_LT; light green), the double Kalman filter (KF_X2; magenta), the combined moving average and lead time Kalman filter post-processing method (MA + KF_LT; red), the 30-day moving average (MA; green), the 30-day linearly weighted moving average (WMA; orange) and the raw numerical weather prediction (NWP) model output (NWP; black).

FIGURE 10  Percentage of forecast busts, defined as absolute forecast errors >3°C, as a function of lead time for all stations for the diurnal bias correction Kalman filter (KF_DBC; blue), the lead time Kalman filter (KF_LT; light green), the double Kalman filter (KF_X2; magenta), the combined moving average and lead time Kalman filter post-processing method (MA + KF_LT; red), the 30-day moving average (MA; green), the 30-day linearly weighted moving average (WMA; orange) and the raw numerical weather prediction (NWP) model output (NWP; black), based on (a) Global Forecast System (GFS) and (b) HARMONIE forecast data.
HARMONIE forecast data; however, all methods reduce the fraction of forecast busts.

The spatial distribution of overall RMSE for each of the six post-processing methods and the raw GFS and HARMONIE forecast data is shown in Figures 11 and 12, respectively. Each circle indicates the location of the station, and the colour represents the overall RMSE, calculated for the whole forecast period, using data from the station at that location. The RMSE for the raw GFS forecast data is in the range 0.9–2.6°C. Generally, the post-processed forecasts show a reduction in RMSE (lighter colours). However, the difference between the raw GFS forecast and the post-processed forecast using KF_LT is less evident here. The RMSE of the raw HARMONIE forecast data is generally smaller compared with the raw GFS forecast data, as indicated by the lighter colours. Furthermore, the difference between the post-processed forecasts and the raw HARMONIE forecast is less evident. However, a decrease in RMSE can be seen, mostly for the stations at which the RMSE of the raw HARMONIE forecast data is larger to begin with.

To assess how well the different post-processing methods improve upon the raw NWP forecast for all 100 stations included in the study, the bias (in absolute values), STD and RMSE skill scores were calculated (Equation (17)). The fraction of stations for which post-processing reduced the bias, STD and RMSE by 5% or more (i.e., a skill score of 0.05 or higher) was then calculated as a function of lead time. The results are shown in Tables 1–6. For clarity, and in order to match the results obtained from post-processing of GFS forecasts, the results from the post-processing of HARMONIE forecasts are shown for every third forecast hour.

The results for post-processing of GFS forecasts show that KF_LT reduces the bias, STD and RMSE for the fewest stations. For the very short lead times it reduces the bias, STD and RMSE for almost all stations. However, for the longer lead times, less than 50% of the stations benefit

![Figure 11](image-url)  
**Figure 11** Spatial distribution of RMSE. The circles show the location for each of the 100 stations and the colormap indicates the overall RMSE, calculated for the whole forecast period. The results are based on post-processing of Global Forecast System (GFS) forecast data. The different post-processing methods, as well as the raw numerical weather prediction (NWP) model output are shown in individual subplots; (a) the diurnal bias correction Kalman filter (KF_DBC); (b) the lead time Kalman filter (KF_LT); (c) the double Kalman filter (KF_X2); (d) the combined moving average and lead time Kalman filter post-processing method (MA + KF_LT); (e) the 30-day moving average (MA); (f) the 30-day linearly weighted moving average (WMA); and (g) the raw GFS forecast.
from post-processing using KFLT with respect to bias, whereas less than 20% benefit from the post-processing with respect to STD and RMSE. The other five post-processing methods perform similar to each other. They all reduce the bias and improve the STD of at least 75% of the stations for all lead times. The RMSE values for at least 90% of the stations are improved by at least 5% or more (compared with the raw GFS forecast) for all lead times. Calculating the fraction of stations for which the bias, STD and RMSE increase by 5% or more (not shown) shows that a maximum of 36% of the stations have a bias that is larger than the raw GFS forecast data for a few of the longer lead times when using the KFLT post-processing method. The other five post-processing methods have an increase in bias for maximum 20% of the stations, which can be seen only for the longest lead times. The corresponding results for STD and RMSE show that only 1% of the stations have an STD or RMSE that is 5% or higher compared with the raw GFS forecast, for all post-processing methods.

The corresponding results for the post-processing of HARMONIE forecast data show slightly different results. KFLT is the post-processing method which reduces the bias, STD and RMSE for the fewest stations here as well: less than 9% of the stations for bias and 0% of the stations for STD and RMSE, for some of the longer lead times. The largest difference compared with the results obtained from post-processing GFS forecast data can be seen for how KFDBC and KFX2 perform with respect to bias. Less than 40% of the stations benefit from these post-processing methods for the longer lead times. MA + KFLT, MA and WMA, on the other hand, reduce the bias for more than 90% of the stations for all lead times. The fraction of stations for which the STD is reduced is much smaller than the corresponding fraction for the GFS forecast data. Most of the stations still benefit from the post-processing for the very short lead times; however, only 10% of the stations benefit from the post-processing for the longer lead times. The corresponding results based on

**FIGURE 12** Spatial distribution of RMSE. The circles show the location for each of the 100 stations and the colormap indicates the overall RMSE, calculated for the whole forecast period. The results are based on post-processing of HARMONIE forecast data. The different post-processing methods, as well as the raw numerical weather prediction model output, are shown in individual subplots; (a) the diurnal bias correction Kalman filter (KFDBC); (b) the lead time Kalman filter (KFLT); (c) the double Kalman filter (KFX2); (d) the combined moving average and lead time Kalman filter post-processing method (MA + KFLT); (e) the 30-day moving average (MA); (f) the 30-day linearly weighted moving average (WMA); and (g) the raw HARMONIE forecast.
### TABLE 1
Percentage of weather stations improved by 5% or more from the post-processing of GFS forecast data for each lead time for each of the six post-processing (PP) methods tested. The results are based on the percentage of stations for which the abs(bias) skill score was 5% or higher.

<table>
<thead>
<tr>
<th>PP method</th>
<th>+3 h</th>
<th>+6 h</th>
<th>+9 h</th>
<th>+12 h</th>
<th>+15 h</th>
<th>+18 h</th>
<th>+21 h</th>
<th>+24 h</th>
<th>+27 h</th>
<th>+30 h</th>
<th>+33 h</th>
<th>+36 h</th>
<th>+39 h</th>
<th>+42 h</th>
<th>+45 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFDBC</td>
<td>96</td>
<td>92</td>
<td>95</td>
<td>91</td>
<td>94</td>
<td>90</td>
<td>91</td>
<td>83</td>
<td>81</td>
<td>79</td>
<td>81</td>
<td>79</td>
<td>74</td>
<td>76</td>
<td></td>
</tr>
<tr>
<td>KFLT</td>
<td>96</td>
<td>98</td>
<td>88</td>
<td>79</td>
<td>64</td>
<td>71</td>
<td>80</td>
<td>79</td>
<td>82</td>
<td>62</td>
<td>37</td>
<td>27</td>
<td>24</td>
<td>37</td>
<td>67</td>
</tr>
<tr>
<td>KFX2</td>
<td>99</td>
<td>93</td>
<td>94</td>
<td>90</td>
<td>95</td>
<td>90</td>
<td>96</td>
<td>85</td>
<td>83</td>
<td>79</td>
<td>80</td>
<td>75</td>
<td>79</td>
<td>75</td>
<td>80</td>
</tr>
<tr>
<td>MA + KFLT</td>
<td>97</td>
<td>89</td>
<td>91</td>
<td>87</td>
<td>92</td>
<td>83</td>
<td>90</td>
<td>84</td>
<td>85</td>
<td>82</td>
<td>81</td>
<td>80</td>
<td>83</td>
<td>81</td>
<td>85</td>
</tr>
<tr>
<td>MA</td>
<td>91</td>
<td>87</td>
<td>92</td>
<td>85</td>
<td>92</td>
<td>86</td>
<td>82</td>
<td>83</td>
<td>82</td>
<td>85</td>
<td>80</td>
<td>83</td>
<td>80</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>WMA</td>
<td>94</td>
<td>89</td>
<td>93</td>
<td>89</td>
<td>93</td>
<td>84</td>
<td>93</td>
<td>84</td>
<td>86</td>
<td>82</td>
<td>85</td>
<td>82</td>
<td>87</td>
<td>82</td>
<td>87</td>
</tr>
</tbody>
</table>

### TABLE 2
Percentage of weather stations improved by 5% or more from the post-processing of GFS forecast data for each lead time for each of the six post-processing (PP) methods tested. The results are based on the percentage of stations for which the STD skill score was 5% or higher.

<table>
<thead>
<tr>
<th>PP method</th>
<th>+3 h</th>
<th>+6 h</th>
<th>+9 h</th>
<th>+12 h</th>
<th>+15 h</th>
<th>+18 h</th>
<th>+21 h</th>
<th>+24 h</th>
<th>+27 h</th>
<th>+30 h</th>
<th>+33 h</th>
<th>+36 h</th>
<th>+39 h</th>
<th>+42 h</th>
<th>+45 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFDBC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>95</td>
<td>96</td>
<td>92</td>
<td>97</td>
</tr>
<tr>
<td>KFLT</td>
<td>100</td>
<td>100</td>
<td>70</td>
<td>36</td>
<td>36</td>
<td>32</td>
<td>66</td>
<td>79</td>
<td>58</td>
<td>22</td>
<td>19</td>
<td>11</td>
<td>13</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>KFX2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>90</td>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>MA + KFLT</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>99</td>
<td>97</td>
<td>99</td>
<td>96</td>
<td>97</td>
<td>93</td>
<td>97</td>
<td>90</td>
<td>96</td>
<td>89</td>
</tr>
<tr>
<td>MA</td>
<td>99</td>
<td>98</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>99</td>
<td>96</td>
<td>98</td>
<td>94</td>
<td>97</td>
<td>92</td>
<td>97</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>WMA</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>94</td>
<td>97</td>
<td>94</td>
<td>97</td>
<td>92</td>
<td>97</td>
<td>92</td>
</tr>
</tbody>
</table>

### TABLE 3
Percentage of weather stations improved by 5% or more from the post-processing of GFS forecast data for each lead time for each of the six post-processing (PP) methods tested. The results are based on the percentage of stations for which the RMSE skill score was 5% or higher.

<table>
<thead>
<tr>
<th>PP method</th>
<th>+3 h</th>
<th>+6 h</th>
<th>+9 h</th>
<th>+12 h</th>
<th>+15 h</th>
<th>+18 h</th>
<th>+21 h</th>
<th>+24 h</th>
<th>+27 h</th>
<th>+30 h</th>
<th>+33 h</th>
<th>+36 h</th>
<th>+39 h</th>
<th>+42 h</th>
<th>+45 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFDBC</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>99</td>
<td>98</td>
<td>95</td>
<td>96</td>
<td>92</td>
<td>97</td>
</tr>
<tr>
<td>KFLT</td>
<td>100</td>
<td>100</td>
<td>70</td>
<td>36</td>
<td>36</td>
<td>32</td>
<td>66</td>
<td>79</td>
<td>58</td>
<td>22</td>
<td>19</td>
<td>11</td>
<td>13</td>
<td>9</td>
<td>28</td>
</tr>
<tr>
<td>KFX2</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>97</td>
<td>93</td>
<td>96</td>
<td>90</td>
<td>96</td>
<td>90</td>
</tr>
<tr>
<td>MA + KFLT</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>98</td>
<td>99</td>
<td>97</td>
<td>99</td>
<td>96</td>
<td>97</td>
<td>93</td>
<td>97</td>
<td>90</td>
<td>96</td>
<td>89</td>
</tr>
<tr>
<td>MA</td>
<td>99</td>
<td>98</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>97</td>
<td>99</td>
<td>96</td>
<td>98</td>
<td>94</td>
<td>97</td>
<td>92</td>
<td>97</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>WMA</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>94</td>
<td>97</td>
<td>94</td>
<td>97</td>
<td>92</td>
<td>97</td>
<td>92</td>
</tr>
</tbody>
</table>

### TABLE 4
Percentage of weather stations improved by 5% or more from the post-processing of HARMONIE forecast data for each lead time for each of the six post-processing (PP) methods tested. The results are based on the percentage of stations for which the abs(bias) skill score was 5% or higher.

<table>
<thead>
<tr>
<th>PP method</th>
<th>+3 h</th>
<th>+6 h</th>
<th>+9 h</th>
<th>+12 h</th>
<th>+15 h</th>
<th>+18 h</th>
<th>+21 h</th>
<th>+24 h</th>
<th>+27 h</th>
<th>+30 h</th>
<th>+33 h</th>
<th>+36 h</th>
<th>+39 h</th>
<th>+42 h</th>
<th>+45 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>KFDBC</td>
<td>98</td>
<td>100</td>
<td>98</td>
<td>92</td>
<td>89</td>
<td>83</td>
<td>77</td>
<td>69</td>
<td>66</td>
<td>58</td>
<td>53</td>
<td>47</td>
<td>44</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>KFLT</td>
<td>99</td>
<td>92</td>
<td>57</td>
<td>26</td>
<td>22</td>
<td>32</td>
<td>51</td>
<td>70</td>
<td>46</td>
<td>15</td>
<td>8</td>
<td>9</td>
<td>11</td>
<td>20</td>
<td>31</td>
</tr>
<tr>
<td>KFX2</td>
<td>98</td>
<td>100</td>
<td>96</td>
<td>91</td>
<td>89</td>
<td>83</td>
<td>77</td>
<td>69</td>
<td>65</td>
<td>58</td>
<td>53</td>
<td>47</td>
<td>44</td>
<td>37</td>
<td>35</td>
</tr>
<tr>
<td>MA + KFLT</td>
<td>98</td>
<td>98</td>
<td>97</td>
<td>96</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>91</td>
<td>91</td>
<td>91</td>
<td>92</td>
<td>92</td>
</tr>
<tr>
<td>MA</td>
<td>95</td>
<td>96</td>
<td>97</td>
<td>96</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>94</td>
<td>95</td>
<td>91</td>
<td>91</td>
<td>91</td>
<td>92</td>
<td>93</td>
<td>91</td>
</tr>
<tr>
<td>WMA</td>
<td>96</td>
<td>98</td>
<td>98</td>
<td>97</td>
<td>96</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>93</td>
<td>93</td>
<td>92</td>
<td>95</td>
<td>93</td>
<td></td>
</tr>
</tbody>
</table>
RMSE show that more than half of the stations benefit from post-processing for the first forecast day, when using MA+KFLT, MA or WMA and a little less than 40% for the longest lead times. The corresponding percentages for the other post-processing methods are in general lower. The fraction of stations for which the bias, STD and RMSE increase by 5% or more compared with the raw HARMONIE forecast was also calculated (not shown). The bias is the metric showing the largest degradation, with up to 64% of the stations for the longest lead times using KFDBC and KFX2. No stations show an increase in STD, whereas up to 15% of the stations show an increase in RMSE for the longest lead times, using KFDBC and KFX2.

5 | DISCUSSION

The presented results show that all six post-processing methods successfully correct the raw NWP forecasts. Comparing the results from the example stations (Figures 2–7), it can be seen that the lead time Kalman filter, KFLT, reduces the bias, MAE and STD mostly for the very short lead times for both GFS and HARMONIE forecasts, as it was constructed to do. This is an effect from the strong correlation between the current real-time bias and the near-future bias.

The two MAs, MA and WMA, behave similar with respect to bias when applied to GFS and HARMONIE forecast data; they reduce the bias close to zero for all lead times. However, the two MAs hardly reduce the MAE and STD for the HARMONIE forecast data. A reduction in MAE is observed only for Tørring station. For the GFS forecast data, the two MAs can be seen to reduce both the MAE and STD for all four example stations. An explanation for this is that the bias, MAE and STD of the raw GFS forecast data for the four example stations are larger than for the raw HARMONIE forecast data—an expected feature due to the coarser resolution of the GFS forecast. The larger the systematic errors in the raw forecast, the larger improvements can be obtained from post-processing. Hence, the reason for the observed improvement for the GFS forecast and the lack of similar improvement for the HARMONIE forecasts when applying the MAs is a difference in systematic errors for the raw forecasts.

The combined MA and lead time Kalman filter, MA+KFLT, showed improvements for both longer and shorter lead times in the presence of larger systematic errors, and only improvements in the shorter range for smaller systematic errors. The larger improvement for the shorter lead times is attributed to KFFLT, which corrects for the real-time bias, whereas the MA corrects for the systematic biases for the longer lead times.
It could be seen that the bias obtained after post-processing both the GFS and HARMONIE forecast data using the diurnal bias correction Kalman filter, $\text{KF}_{\text{DBC}}$, and the double Kalman filter, $\text{KF}_{\text{X2}}$, follows the shape of the raw NWP forecast. The tendency of $\text{KF}_{\text{DBC}}$ to produce bias results as a function of forecast lead time that follow the shape of the bias for the raw NWP forecast was also noted by Homleid (1995).

From the verification of bias for the example stations it can be seen that the raw HARMONIE model output exhibits a cold drift with lead time (Figure 5). This is a feature that could be seen for the majority of stations. The drift of the HARMONIE forecast bias with lead time of up to about half a degree for individual stations during the 2-day forecast is quite normal, and for the current version of DMI-HARMONIE (v. 40 h1) it has recently mainly been a negative trend (Yang, 2020). A similar cold drift, however not as pronounced, is observed for GFS forecasts for some stations. The identified drifts are with respect to the observed 2 m temperatures as measured by the FieldSense stations, which may not represent the actual “truth”, only the actual temperature recorded.

The percentage of forecast busts (Figure 10) highlights the difference in systematic errors between the raw GFS and HARMONIE forecasts. Overall, the GFS forecasts have a higher fraction of forecast busts compared with the HARMONIE forecasts. It is interesting to note that the post-processing of the GFS forecast data gives comparable fractions of forecast busts to the raw HARMONIE forecasts. This shows that post-processing of a coarse-resolution model can improve the model with respect to large errors so that the performance is comparable to a high-resolution model.

The summary statistics over all 100 stations (Figures 8 and 9) showed that for the GFS forecasts, the two combined post-processing methods, $\text{KF}_{\text{X2}}$ and $\text{MA} + \text{KF}_{\text{LT}}$, are superior. Application of $\text{KF}_{\text{DBC}}$ and $\text{KF}_{\text{X2}}$ tends to result in a reduction in bias for GFS forecast data. The results obtained here are comparable to the results obtained by Doeswijk and Keesman (2005). However, for HARMONIE forecast data, these two post-processing methods tend to increase the bias for the longer lead times. The diurnal bias correction Kalman filter uses the same correction for the first and second forecast days. Therefore, any systematic difference in forecast errors between the first and second forecast days will lead to an inaccurate correction for the last forecast day. This might be the explanation for the unsatisfactory performance observed for $\text{KF}_{\text{DBC}}$ and $\text{KF}_{\text{X2}}$ when applied to HARMONIE forecasts. Furthermore, the systematic errors in the HARMONIE forecasts are expected to be smaller compared with the GFS forecasts. Hence, the linear correlation between the current forecast error and the forecast error at the longer lead times might not be as systematic in the HARMONIE forecast data. This would also affect the corrections and introduce errors with increasing lead time since the structure of the system noise covariance matrix, $Q_k$, is based on the linear correlation between forecast and observed temperature errors. It is instead $\text{MA} + \text{KF}_{\text{LT}}$ that performs the best for HARMONIE forecasts. Overall, seen to post-processing of both GFS and HARMONIE forecasts, $\text{MA} + \text{KF}_{\text{LT}}$ performs the best. It combines the correction of longer-term systematic biases through the moving average with the short-term corrections from the lead time Kalman filter to produce a post-processing method that works best for short-term forecasts but also yields satisfactory results for the whole 2-day forecast.

The results for the example stations highlighted differences in performance among stations, and Figures 11 and 12 showed the spatial distribution of overall RMSE among the stations. Differences in the spatial distribution of RMSE between the raw GFS and HARMONIE forecasts can be seen. The raw GFS forecasts show larger RMSE for stations in northeastern Jutland and also for coastal stations on Zealand. However, larger RMSE can also be seen for stations in central Jutland and no clear relationship between distance to the coast for stations at Fyn can be seen. Overall, there does not seem to be any relationship between larger RMSE and distance to the coast. The results for HARMONIE forecast data also show little correlation between larger RMSE and distance to the coast; both stations close to the coast and stations further inland show larger RMSE.

The weather in Denmark is characterized by prevalent westerly winds, and westerly weather systems are frequent (Cappelen, 2020). However, no relationship between exposure to westerly weather systems and larger systematic errors can be seen here.

Anadranistakis et al. (2004) and Cheng and Steenburgh (2007) evaluated the performance of the Kalman filter, and other post-processing methods, for regions with complex terrain and obtained good performances. The topography for the selected region in this study is much less complex; except for land–sea contrasts Denmark is generally flat and homogeneous. No clear relationship between higher RMSE and topography was found.

Cheng and Steenburgh (2007) compared MOS, moving-average bias removal and the Kalman filter post-processing methods. They showed that the MA and Kalman filter performed better than MOS for quiescent cool season patterns when winter-time cold pools exist, and for quiescent warm season patterns all three methods performed similarly. However, when there are sudden changes in the weather conditions, the MA and Kalman filter do not adapt quickly enough and produce inferior forecasts compared with MOS. To see if there was any difference in performance based on month, the monthly
summary statistics were calculated (not shown). These show that all six post-processing methods have a tendency to increase the STD for the longer lead times for weather-wise vivid winter-time months. Similar results were obtained for the summary statistics of the MAE. Homleid (1995) also found an increase in standard deviation for winter-time months.

In preliminary runs, the system noise covariance $Q$ used in the Kalman filters was time-invariant during the whole period of interest. However, this led to an increase in STD and MAE for GFS forecasts for the later half of the forecast period of interest. This is most probably due the short period that was initially used to estimate the system noise covariance matrix. Furthermore, changing weather regimes likely also affect the correlation between forecast errors for different lead times. Correlations obtained for one season is therefore not likely to be a good fit for another season. However, a sufficiently long observational record does not exist in order to estimate the correlations for different seasons. Hence, the current approach, where the structure of the system noise covariance matrix was updated once every day using forecasts and observations from the last 30 days, was implemented.

Several different methods for how to either choose or estimate the system and observation noise covariance matrices have been suggested. Several studies use time-invariant noise covariances and obtain good results (Doeswijk & Keesman, 2005; Homleid, 1995). However, in doing so, the adjustability of the filter decreases since it will not be able to adjust to changes brought on by, for example, changing weather regimes. Galanis and Anadranistakis (2002) suggest that both the system and observation noise covariance matrices are to be calculated based on the difference between the observed forecast error and the forecast error as estimated from the Kalman filter from the last 7 days. It might be worth investigating if such a method might improve upon the results obtained here. This would also introduce a time-varying aspect to the magnitude of $Q$ and $R$, which would result in different rates of responsiveness to different weather conditions. However, this method is also based on past differences and it is not clear how this would improve the Kalman filter’s response to sudden changes. Another suggestion is to reduce the ratio $Q/R$ with lead time to limit the larger errors that might occur for the longer lead times, especially in situations with rapidly changing weather conditions. In order to obtain a time-varying aspect to the rate of adaptability of the Kalman filter, one could use ensemble forecasts to estimate future uncertainties and base the ratio $Q/R$ on this. Furthermore, analogue forecasting could be used to classify the weather situations that give rise to the largest errors and this could then be used to adjust the noise covariances for these situations. However, analogue forecasting requires a long historical archive of past forecasts and observations and would therefore be contradictory to the reason for choosing the Kalman filter, which was to use a post-processing method that does not require a long historical archive.

All six post-processing methods tested in this study are linear. Non-linear post-processing methods based on machine learning, such as neural networks and random forests, have shown good results when used to post-process NWP forecasts (Casaioi et al., 2003; Eccel et al., 2007; Marzban, 2003). The advantages of these methods are that they can represent non-linear relationships and that several predictors can be used easily. However, these post-processing methods require a long training dataset for optimal performance. Post-processing methods that require a long historical archive are challenging and computationally expensive in an operational set-up since the state-of-the-art NWP models are frequently updated, which requires re-forecasts to be made to obtain these records for post-processing applications.

6 | CONCLUSIONS

In this paper, we have compared the performance of six post-processing methods based on results from application to 2 m temperature forecasts from the relatively coarse-resolution global numerical weather prediction (NWP) model Global Forecast System (GFS) and the regional high-resolution NWP model HARMONIE. Both post-processing methods handling the longer-term systematic errors and post-processing methods focusing on the shorter-term systematic errors were evaluated. Compared to other studies, which have focused on the post-processing of 2 m temperature forecasts in areas with more complex topography (Anadranistikis et al., 2004; Cheng & Steenburgh, 2007; Crochet, 2004), this study has focused on Denmark, which is a region with a relatively homogeneous topography. Here, no clear relationship between larger systematic errors and topography was found. Furthermore, no clear connections between the distance to the coast and larger systematic errors were. In addition, no connections between stations more prone to exposure to westerly weather systems and larger systematic errors were found.

It was shown that the performance of most of the post-processing methods were similar for the two NWP forecasts. However, the performance of the diurnal bias correction Kalman filter and the double Kalman filter differed. Both of these post-processing methods have been successfully applied to 2 m temperature forecasts from coarser-resolution NWP models before in study areas with both relatively homogeneous and complex topography,
and have yielded good results (Doeswijk & Keesman, 2005; Homleid, 1995). Similarly, in this study, both of these post-processing methods resulted in an improvement of the 2m temperature GFS forecast. However, we also showed that for a high-resolution regional NWP model for a region with relatively homogeneous topography, the diurnal bias correction Kalman filter and the double Kalman filter do not perform satisfactorily. They performed worse for the longer lead times with larger mean absolute errors compared with the raw HARMONIE forecasts. The reason for this is believed to be a combination of the fact that the diurnal bias correction Kalman filter applies the same correction to the first and second forecast days, and that the errors for the HARMONIE forecast data are generally smaller and not as systematic as for the GFS forecast data.

The 30-day moving average followed by the lead time Kalman filter was shown to perform the best with respect to both GFS and HARMONIE forecasts. The moving average was shown to be superior to the diurnal bias correction Kalman filter at removing the longer-term systematic errors for HARMONIE forecast data and comparable for GFS forecast data. Subsequent application of the lead time Kalman filter corrects for the short-term errors using the real-time forecast error. The post-processing method is adaptive and the structure of the system noise covariance matrix is updated using the last 30 days of forecast–observation pairs. Therefore, there is no need for a long record of observations or a historical archive of forecasts to implement the method. The combined moving average and lead time Kalman filter post-processing method reduces the bias of the GFS and HARMONIE forecasts close to zero for all forecast lead times and also reduces the standard deviation and RMSE of the two forecasts for the majority of stations. This shows that the application of a relatively simple post-processing method can give good results.

ACKNOWLEDGEMENTS
This study was partially funded by Innovation Fund Denmark, through grant 8053-00242B. The authors wish to thank FieldSense for providing observational data from their network of private weather stations and DMI for providing weather forecast data from the DMI-HARMONIE NWP model.

AUTHOR CONTRIBUTIONS
Emy Alerskans: Conceptualization (equal); formal analysis (lead); investigation (lead); methodology (lead); visualization (lead); writing – original draft (supporting); writing – review and editing (equal).
Eigil Kaas: Conceptualization (equal); formal analysis (supporting); investigation (supporting); methodology (supporting); visualization (supporting); writing – original draft (supporting); writing – review and editing (equal).

CONFLICT OF INTEREST
The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT
Observational data were provided by FieldSense from their network of private weather stations. Numerical weather prediction data from the HARMONIE model were provided by the Danish Meteorological Institute. Due to their proprietary nature, these data cannot be made openly available. Numerical weather prediction data from the GFS model are available from the National Center for Atmospheric Research (NCAR) Research Data Archive (https://rda.ucar.edu/).

ORCID
Emy Alerskans https://orcid.org/0000-0001-8206-5999
Eigil Kaas https://orcid.org/0000-0001-6970-2404

REFERENCES


Klein, W.H., Lewis, B.M. & Enger, I. (1959) Objective prediction of five-day mean temperatures during winter. *Journal of Atmospheric Sciences*, 16(6), 672–682.


Yang, X. (Personal communication, 9th September 2020).

Yang, X., Andersen, B., Dahlbom, M., Sass, B., Zhuang, S., Amstrup, B. et al. (2017) NEA, the operational implementation of HARMONIE 40h1.1 at DMI. *ALADIN–HIRLAM Newsletter*, 8, 104–111.


**How to cite this article:** Alerskans, E., & Kaas, E. (2021). Local temperature forecasts based on statistical post-processing of numerical weather prediction data. *Meteorological Applications*, 28(4), e2006. [https://doi.org/10.1002/met.2006](https://doi.org/10.1002/met.2006)