Improving Short-term Forecasts of Daily Maximum Temperature with the Kalman Filter with GMM Estimation*

Marco Costa[1,4][0000−0001−7686−2430], Fernanda Catarina Pereira[2,5][0000−0002−6545−2900], and A. Manuela Gonçalves[3,5][0000−0001−8491−6048]

[1] University of Aveiro, Águeda School of Technology and Management - ESTGA, Portugal. marco@ua.pt
[2] University of Minho, Department of Mathematics, Portugal. up202010700@edu.fe.up.pt
[3] University of Minho, Department of Mathematics, Portugal. mneves@math.uminho.pt
[4] Centre for Research and Development in Mathematics and Applications - CIDMA, University of Aveiro, Portugal
[5] Center of Mathematics, University of Minho, Portugal

Abstract. Within the scope of the TO CHAIR project, a state space modeling approach is proposed in order to improve accuracy obtained from the *weatherstack.com* website with a dataset of real observations. The proposed model establishes a stochastic linear relationship between the maximum temperature observed and the h-step-ahead forecast produced from the website. This relation is modeled in a state space framework associated to the Kalman filter predictors. Since normality of disturbances was not a good assumption for this dataset, alternative Generalized Method of Moments (GMM) estimators were considered in the models parameters estimation. The results show that this approach allows reducing the RMSE of the uncorrected forecasts in 16.90% considering the 6-step-ahead forecasts and in 60.45% considering the 1-step-ahead forecasts, compared with the initial RMSE. Additionally, empirical confidence intervals at the 95% level have a coverage rate similar to this confidence level. So, this approach has proven suitable for this type of forecasts correction since it considers a stochastic calibration factor in order to model time correlation of this type of variable.

Keywords: State space modeling · Kalman filter · GMM estimation · Forecasting calibration · Maximum temperature · TO CHAIR project

* This work has received funding from FEDER/COMPETE/NORTE2020/POCI/FCT funds through grants UID/EEA/- 00147/20 13/UID/IEEA/00147/ 006933-SYSTEC, project and To CHAIR - POCI-01-0145-FEDER-028247. This work was also partially supported by the Portuguese FCT Projects UIDB/00013/2020 and UIDP/00013/2020 of CMAT-UM and the Center for Research and Development in Mathematics and Applications (CIDMA) through the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), references UIDB/04106/2020 and UIDP/04106/2020.

1 Introduction

This work is carried out in the context of project "TO CHAIR - Optimum Challenges in Irrigation" - https://systec.fe.up.pt/projects/FCT-TOCHAIR/ - and aims to understand and analyze the behavior of humidity in the soil by mathematical/statistical modeling in order to find optimal solutions to improve the efficiency of daily water use in irrigation systems [3].

In the context of the TO CHAIR project, it is necessary to improve short-term forecasts of meteorological variables. In fact, more accurate forecasts of these variables can improve the results of the optimization routines in order to obtain a more efficient use of water in irrigation systems.

In this project, the main goal of statistical modeling is to improve the accuracy of the forecast of meteorological variables obtained from the *weatherstack.com* website for the location under analysis, a farm in Portugal. However, agricultural researchers that investigate in this area know that forecasts have significant errors compared with observations obtained locally by a portable weather station. Several factors can contribute to these discrepancies. On the one hand, this farm is located in a valley in a mountainous region, and so it has a specific orography. On the other hand, the methodology adopted by the site's forecasts (which we do not know), possibly associated with the significant distance between this farm and fixed weather stations in with which forecasts are computed, can partially explain these differences.

This work intends to establish a state space framework that combines forecasts with the observations in order to correct or "calibrate" a forecast by comparing it with the knowledge from the past, namely through an estimated model based on few data. This approach has been considered in environmental problems, for instance in [1, 4].

2 Exploratory Analysis of Data

The statistical analysis was performed using a dataset that includes forecasts (obtained from the weatherstack.com website) of daily maximum temperature (in Celsius degrees) for the location of the farm Senhora da Ribeira in Portugal, between February 20 and October 11, 2019. Additionally, we also use observations of daily temperatures obtained by a portable weather station installed in the farm during that period of 234 days (see Fig. 1).

In this context, we consider that Y_t is the real maximum temperature in day t with a small error associated to the measurement of the portable station. However, the forecast $W_{t:t-h}$ has an additional uncertainty associated to the interpolation methods or the methodology adopted by the website.

The data from the portable station will be used to compare with the site's forecasts and to assess their accuracy. Considering that the observations from the portable weather station are more accurate, in fact, the most accurate observations available, they will be used for correct or calibrate the site's forecasts, since the portable station was temporary installed in the farm.

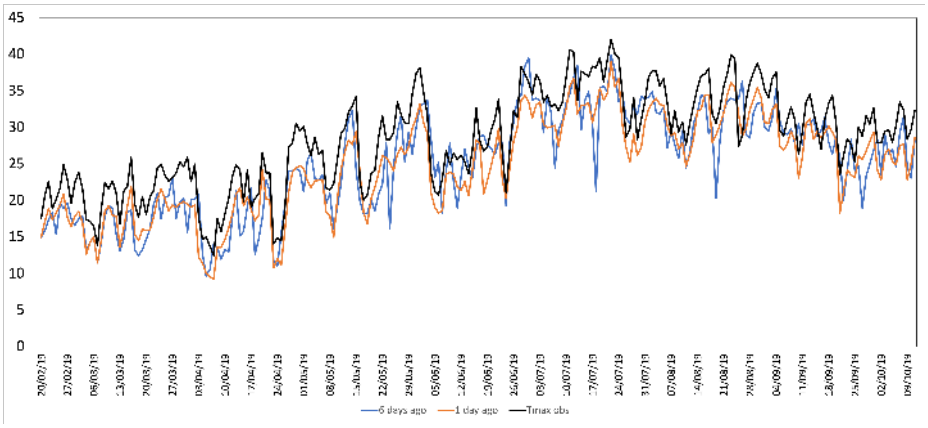

Fig. 1. Time series plots of the observed maximum temperature (in black) and the respective forecasts at 6-steps-ahead and 1-step-ahead (in blue and orange, respectively).

The roots of the mean square error of the forecasts from the site, $W_{t:t-h}$, with $t = 1, ..., 234$ and $h = 1, ..., 6$, compared with the observations Y_t, with $t = 1, ..., 234$, computed by

$$\text{RMSE}_h = \sqrt{\frac{1}{234}\sum_{t=1}^{234}\left(Y_t - W_{t:t-h}\right)^2} \tag{1}$$

where $W_{t:t-h}$ represents the h-steps-ahead forecast of the maximum temperature in day t, that is, the forecast indicated by the site h days before the day t, and Y_t is the observed maximum temperature in the farm by the portable weather station.

Table 1 presents the root mean square error, RMSE_h. Notice that, as expected, the RMSE is greater for large values of h than for forecasts obtained with few days of delay. However, all RMSE are very significant, even for forecasts obtained a day before. So, the site's forecasts are significantly inaccurate when compared with the observations collected in the farm.

Table 1. Root of the mean square error (RMSE) between the maximum temperature observed in the farm and the h-steps-ahead obtained from the site weatherstack.com, with $h = 1, ..., 6$.

h-step-ahead	6	5	4	3	2	1
RMSE	4.670	4.222	4.107	4.003	3.901	3.875

However, in spite of the website's inaccurate forecasts, forecasts and observations are linear correlated. In fact, the Pearson's correlation coefficients between

observations from the portable station and the h-step-ahead forecasts show a significant linear correlation (Table 2)

Table 2. Pearson's correlation coefficients between observations from the portable station and the h-step-ahead forecasts show, with $h = 1, ..., 6$.

h-step-ahead	6	5	4	3	2	1
correlation	0.880	0.918	0.941	0.956	0.970	0.976

3 The State Space Approach

3.1 The State Space Model

Considering that the forecasts $W_{t:t-h}$, with $t = 1, ..., 234$, are known at instant $t - h$, and the observed maximum temperature Y_t is related with forecasts, we propose a state space model composed by these two equations:

$$Y_t = \beta_t W_{t:t-h} + e_t \tag{2}$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \epsilon_t \tag{3}$$

where Eq. 2 is the observation equation and Eq. 3 is the state or transition equation. This model assumes that the maximum temperature observed in the farm at day t is linear related with the h-step-ahead forecast given by the website at day $t - h$.

The unobservable process $\{\beta_t\}$ is called the state process and must be predicted. In this case, it is be assumed that the process $\{\beta_t\}$ follows a stationary autoregressive process of order 1, that is, $\{\beta_t\} \sim \text{AR}(1)$, with mean μ and the autoregressive coefficient ϕ, such as $|\phi| < 1$.

Errors e_t and ϵ_t are assumed to be a sequence of uncorrelated variable with zero mean and variances σ_e^2 and σ_ϵ^2, respectively, and uncorrelated with each other, that is, $E(e_t\epsilon_r) = 0, \forall t, r$.

Usually, in several applications it is assumed that the disturbances e_t and ϵ_t are normally distributed, that is, $e_t \sim N(0, \sigma_e^2)$ and $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, however, this assumption is not always valid with environmental data.

The model Eq 2 - Eq 3 assumes that the state process represents a stochastic calibration factor between the maximum temperature observation and the website's forecasts, which contain significant uncertainty. As the factor β_t is stochastic, it varies over time allowing some flexibility in the correction procedure.

3.2 Kalman Filter

The Kalman filter, proposed by Kalman (1960) and Kalman and Bucy (1961), is an iterative algorithm that produces, at each time t, an estimator of the state

vector at time t. It provides optimal unbiased linear one-step-ahead and update estimators of the unobservable state β_t.

Let $\hat{\beta}_{t|t-1}$ denote the predictor of β_t based on the observations $Y_1, Y_2, \ldots, Y_{t-1}$ and $P_{t|t-1}$ be its mean square error (MSE), this is, $E[(\hat{\beta}_{t|t-1} - \beta_t)^2]$. The one-step-ahead forecast for the observable vector Y_t is given by $\hat{Y}_{t|t-1} = W_{t:t-h}\hat{\beta}_{t|t-1}$.

When, at time t, Y_t is available, the prediction error or innovation, $\eta_t = Y_t - \hat{\beta}_{t|t-1}$, is used to update the estimate of β_t (filtering) through the equation

$$\hat{\beta}_{t|t} = \hat{\beta}_{t|t-1} + K_t\eta_t, \tag{4}$$

where K_t is called the Kalman gain matrix and is given by

$$K_t = P_{t|t-1}W_{t:t-h}(W^2_{t:t-h}P_{t|t-1} + \sigma_e^2)^{-1}. \tag{5}$$

Furthermore, the MSE of the updated estimator $\hat{\beta}_{t|t}$, represented by $P_{t|t}$, verifies the relationship $P_{t|t} = P_{t|t-1} - K_tW_{t:t-h}P_{t|t-1}$.

The Kalman filter algorithm is initialized with $\hat{\beta}_{1|0}$ and $P_{1|0}$, and when the state process is stationary, it can be initialized considering that initial state vector β_1 has $\hat{\beta}_{1|0} = \mu$ and MSE $\sigma_\varepsilon^2(1-\phi^2)^{-1}$.

3.3 Parameters Estimation – a Distribution-free Approach

When disturbances e_t and ϵ_t are normally distributed and under the independence between errors and the initial state β_1, the parameters $\Theta = (\mu, \phi, \sigma_e^2, \sigma_\varepsilon^2)$ can be estimated by the Gaussian maximum likelihood method.

The log-likelihood of a sample $(Y_1, Y_2, \ldots, Y_n)$ can be written through conditional distributions, given by

$$\log L(\Theta; Y_1, Y_2, \ldots, Y_n) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{n}\log(\Omega_t) - \frac{1}{2}\sum_{t=1}^{n}\eta_t^2\Omega_t^{-1}, \tag{6}$$

where

$$\Omega_t = W^2_{t:t-h}P_{t|t-1} + \sigma_e^2. \tag{7}$$

The optimization of the log-likelihood is done by numerical procedures via the Newton-Raphson method or, more often, by the EM algorithm ([6]).

However, previous modeling has shown that the normality is rejected in the residuals analysis. So, alternative methods are needed. In this context, we proposed to adapted the distribution-free estimators initially proposed in [2] and subsequently generalize them for multivariate models in [5].

Considering the model of type Eq 2 - Eq 3 for some h, the mean, μ, of the state $\{\beta_t\}_{t=1,2,\ldots}$, can be easily estimated by the generalized method of moments (GMM):

$$\widehat{\mu} = \frac{1}{n}\sum_{t=1}^{n}\frac{Y_t}{W_{t:t-h}}. \tag{8}$$

The autoregressive parameter ϕ is estimated by the covariance structure of process $\{Y_t W_{t:t-h}^{-1}\}_{t=1,2,...}$ based on the autocovariance function of the process $\{\beta_t\}$ by

$$\widehat{\phi} = \frac{\sum_{k=1}^{\ell} \gamma(k+1)\gamma(k)}{\sum_{k=1}^{\ell} \gamma^2(k)} \tag{9}$$

where $\widehat{\gamma}(k)$ is the sample autocovariance function of the process $\{Y_t W_{t:t-h}^{-1}\}_{t=1,2,...}$.

The choice of ℓ was discussed in the original work [2] and for a sample of dimension of 200, as it is approximately in this case, it is recommended the use of $\ell = 60$.

To estimate σ_ϵ^2 it is considered the distribution-free estimator

$$\widehat{\sigma}_\epsilon^2 = \frac{1-\widehat{\phi}^2}{\widehat{\phi}}\widehat{\gamma}(1). \tag{10}$$

The observation noise variance σ_e^2 is based on sample mean square error of the process $\{Y_t W_{t:t-h}^{-1}\}_{t=1,2,...}$, that is, $\widehat{\gamma}_0$, defining $\widehat{\gamma}_0$ as

$$\widehat{\gamma}_0 = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{Y_t}{W_{t:t-h}} - \widehat{\mu}\right)^2$$

and the estimator of σ_e^2 is given by

$$\widehat{\sigma}_e^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{Y_t}{W_{t:t-h}} - \widehat{\mu}\right)^2 - \frac{\widehat{\sigma}_\epsilon^2}{(1-\widehat{\phi}^2)^2}. \tag{11}$$

The estimators 8 to 11 are, under simple regularity conditions, consistent ([2]). In this sense, in this work the dataset has a high sample dimension (up to 200) which ensures the properties of the estimators.

3.4 Forecasts Correction Procedure

Once modeled the relation between the observed maximum temperature and its forecasts h-step-ahead from the website, the correction procedure for them can then be proposed.

At each day t it is available the observation Y_t and six forecasts $W_{t+1:t}$, $W_{t+2:t}$, ..., $W_{t+h:t}$, ..., $W_{t+6:t}$. The main goal is to improve these forecasts with the observations available until the present, the day t.

So, for each h-step-ahead forecast and for each day t it is possible to predict the correction factor β_{t+h} to the day $t+h$ using the Kalman filter prediction, such as

$$\widehat{\beta}_{t+h|t} = \mu + \phi^h(\widehat{\beta}_{t|t} - \mu) \tag{12}$$

with the mean square error

$$P_{t+h|t} = \phi^{2h}P_{t|t} + \phi^{2(h-1)}\sigma_\epsilon^2 + \phi^{2(h-2)}\sigma_\epsilon^2 + ... + \phi^2\sigma_\epsilon^2 + \sigma_\epsilon^2. \tag{13}$$

Thus, the corrected h-step-ahead forecast for Y_{t+h} based on data available until t is given by

$$\widehat{Y}_{t+h|t} = \widehat{\beta}_{t+h|t} W_{t+h:t} \tag{14}$$

with mean square error given by

$$\text{MSE}_{t+h|t} = W^2_{t+h:t} P_{t+h|t} + \sigma^2_e. \tag{15}$$

Even considering the distribution-free estimators, but attending to their asymptotic properties of consistence, we can compute empirical confidence intervals with $(1-\alpha) * 100\%$ level to corrected forecasts from equations

$$Y_{t+h|t} = \widehat{Y}_{t+h|t} \pm z_{1-\frac{\alpha}{2}} \sqrt{\text{MSE}_{t+h|t}}. \tag{16}$$

where $z_{1-\alpha/2}$ is the normal quantil of probability $1-\alpha/2$.

4 Results

The modeling procedure considered two series: the training series (in-sample) and the out-sample series. The first time series comprising data between 20 February until 31 August, 2020 (193 days) is considered to estimate parameters and analyze the assumptions based on the residuals analysis; the second time series comprising data between 1 September and 11 October, 2020 (41 days), is considered to better assess the model's performance in a independent period where the models were adjusted.

Table 3. Estimates of parameters.

parameter / h	6	5	4	3	2	1
μ	1.1650	1.1555	1.1633	1.1608	1.1658	1.1723
ϕ	0.8394	0.7857	0.8092	0.8796	0.9419	0.8888
σ^2_ϵ	0.0028	0.0032	0.0032	0.0008	0.0009	0.0011
σ^2_e	9.3464	6.0780	4.2854	3.8355	2.0164	2.2936

Table 3 presents the estimates of parameters for all six models considering $h = 1, ..., 6$. As expected, these results show that the mean of the state $\{\beta_t\}$ is greater than 1 in all models. This means that, in average, the forecasts are lower than the observations from the portable station; that is, there is a stochastic bias. However, as the state process is a factor, this bias can be interpreted in a relative way. For instance, $\widehat{\mu} = 1.1723$ for the 1-step-ahead forecasts, that is, in average, the maximum temperature observed in the farm was greater in 17.23% than the 1-step-ahead forecast. This factor does not differ much for the different values of h.

All estimates for the autoregressive parameters are less than 1, so, the state process is estimated as a stationary process, as assumed in assumptions. The

error of the observation equation has a high variance in all models. However, this variability decrease as the forecasts are computed with less delay; for the 6-step-ahead forecasts the observation error was estimated in 9.3464 instead of 2.2936 for the 1-step-ahead forecasts. The error of the state equation has a low variability for all h.

In order to verify the assumptions of model, an analysis of residuals $\widehat{\eta}_{t|t-1}$ was performed. The residuals present an uncorrelated structure compatible with assumptions. Fig. 2 represents histograms of the model's residuals for $h = 1$ and $h = 6$.

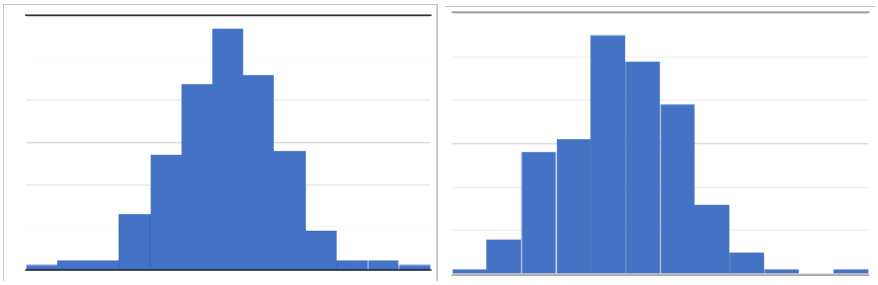

Fig. 2. Histograms of model's residuals $\widehat{\eta}_{t|t-1}$ of models for $h = 1$ and $h = 6$, for whole period under analysis.

The estimates of parameters, associated to the Kalman filter equations, allow to predict the state values, that is, the calibration factors. For instance, Fig. 3 shows the Kalman filter h-step-ahead forecasts of the calibration factors, $\widehat{\beta}_{t+h|t}$ to the model with $h = 1$ and $h = 6$. These predictions show that the calibration factor varies over time, thus showing its variability.

This approach, associated to the Kalman filter, allows to compute h-step-ahead predictions of Y_t based on Eq. 14. Figure 4 shows the observed maximum temperature in the whole period under analysis with the website's forecasts for $h = 1, 6$ and the empirical confidence intervals of the corrected forecasts by the Kalman filter.

For each h days in advance it was computed the root of the mean square error of the corrected forecasts. The RMSE of the corrected forecasts given by the Kalman filter reduced in 16.90% considering the 6-step-ahead forecasts and in 60.45% considering the 1-step-ahead forecasts, compared with the initial RMSE (Table 4).

These results show a big reduction in MSE considering 1 day ahead and lower reduction when h increases, as expect. Moreover, the confidence intervals with a level of 95% have a coverage rate similar to this level (Table 5). Thus, these models are well adjusted and suitable to model this type of data.

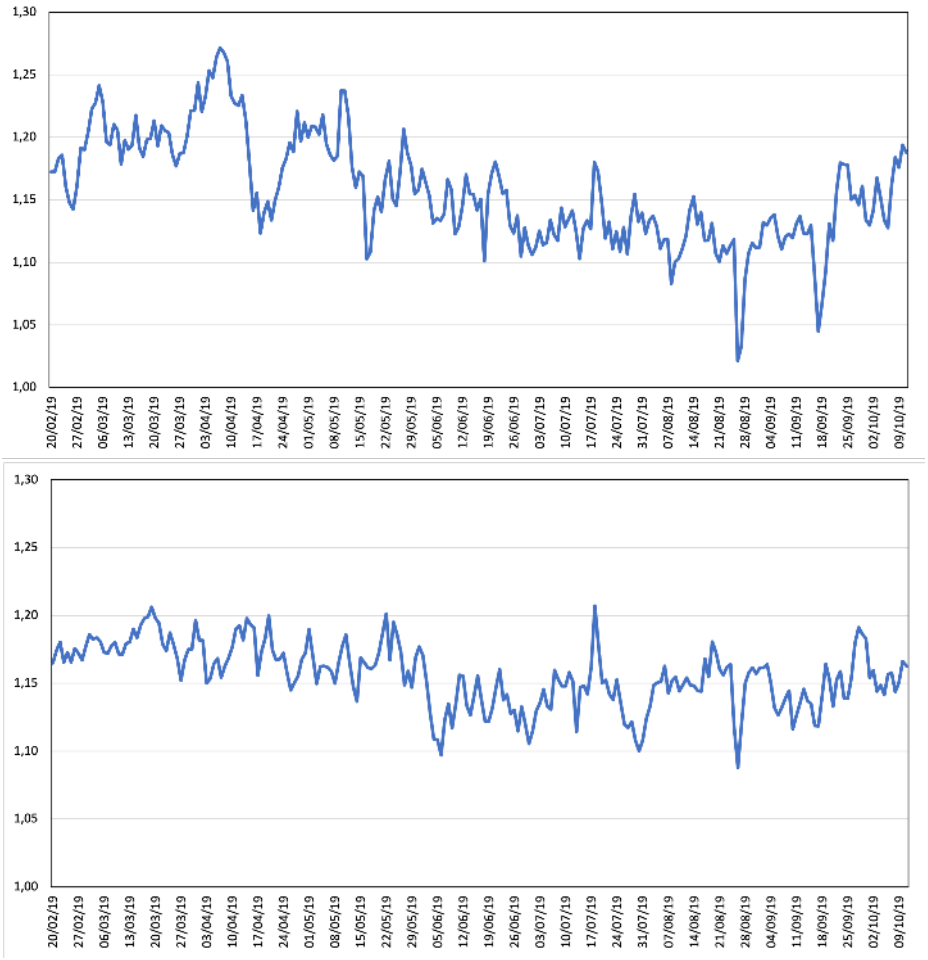

Fig. 3. Kalman filter forecasts of the calibration factor $\widehat{\beta}_{t+h|t}$ for the models with $h = 1$ (up) and $h = 6$ (down).

Table 4. Roots of the mean square errors before and after correction.

h	6	5	4	3	2	1
RMSE in-sample corrected	3.937	3.206	2.826	2.264	1.890	1.545
RMSE out-sample corrected	3.602	2.283	1.737	1.911	1.687	1.469
global RMSE uncorrected	4.670	4.222	4.107	4.003	3.901	3.875
global RMSE corrected	3.881	3.065	2.668	2.206	1.856	1.532
reduction of global RMSE (%)	16.90%	27.41%	35.04%	44.89%	52.42%	60.45%

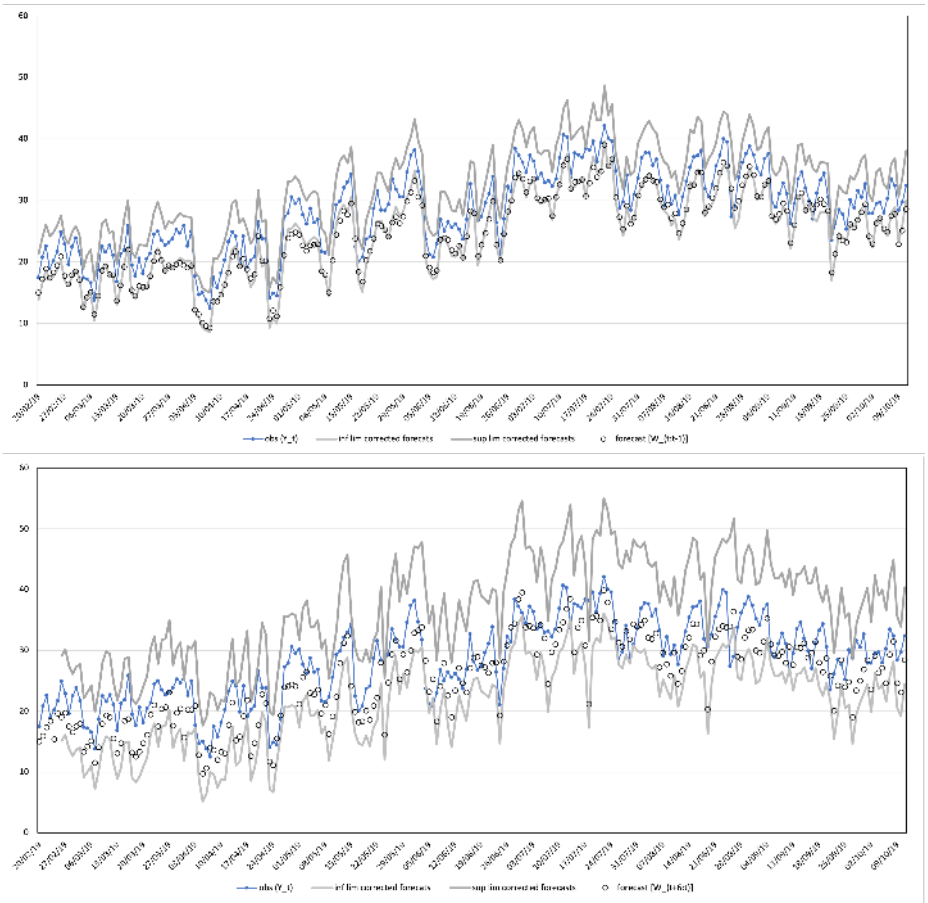

Fig. 4. Empirical confidence intervals for h-step-ahead corrected forecasts of the maximum temperature for model $h = 1$ and $h = 6$, for the whole period under analysis.

Table 5. Coverage rates of empirical confidence intervals of corrected forecasts , with $h = 1, ..., 6$.

h-step-ahead	6	5	4	3	2	1
coverage rate	97.86%	93.97%	95.24%	95.65%	97.38%	95.18%

5 Conclusions

The state space approach shows that it can be considered in the improvement of weather variables forecasts obtained from some accessible sources, even if those sources produce data with a significant errors, as long as, more accurate data is available in order to estimate the parameters model.

Furthermore, as the normality of disturbances was not validated in a previous analysis, the option for distribution-free estimators based on the GMM methods proved to be adequate and produces good reductions in the RMSE of initial forecasts. These reductions were more significant as the delay of the forecasts were smaller.

References

1. Bruno, F., Cocchi, D., Greco, F. et al.: Spatial reconstruction of rainfall fields from rain gauge and radar data. Stoch Environ Res Risk Assess **28**: 1235–1245 (2014). https://doi.org/10.1007
2. Costa, M., Alpuim, T.: Parameter estimation of state space models for univariate observations, Journal of Statistical Planning and Inference, **140**: 1889–1902 (2010)
3. Costa, C., Goncalves, A.M., Costa, M., Lopes, S.: Forecasting Temperature Time Series for Irrigation Planning Problems. In: Proceedings of the 34th IWSM International Workshop on Statistical Modelling, (2019), http://hdl.handle.net/10773/26437
4. Costa, M. and Alpuim, T.: Adjustment of state space models in view of area rainfall estimation. Environmetrics, **22**: 530-540 (2011) https://doi.org/10.1002/env.1064
5. Gonçalves, A.M., Costa, M.: Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. Stoch Environ Res Risk Assess **27** 1021–1038 (2013). https://doi.org/10.1007/s00477-012-0640-7
6. Shumway, R.H., Stoffer, D.F.: *Time series analysis and its applications: with R examples*, Springer, New York, 2011.