MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Memory-Based Modeling of Seasonality for Prediction of Climatic Time Series

Daniel Nikovski, Ganesan Ramachandran

TR2009-050 September 2009

Abstract

The paper describes a method for predicting climate time series that consist of significant annual and diurnal seasonal components and a short-term stockastic component. A memory-based method for modeling of the non-linear seasonal components is proposed that allows the application of simpler linear models for predicting short-term deviations from seasonal averages. The proposed method results in significant reduction of prediction error when predicting error time series of ambient air temperature from multiple locations. Moreover, combining the statistical predictor with meteorological forecasts using linear regression or Kalman filtering further reduces error to typically between $1\,^{\circ}\mathrm{C}$ over a prediction horizon of one hour and $2.5\,^{\circ}\mathrm{C}$ over 24 hours.

Machine Learning Journal

Memory-Based Modeling of Seasonality for Prediction of Climatic Time Series

Daniel Nikovski[1] and Ganesan Ramachandran[2]

[1] Mitsubishi Electric Research Laboratories, Cambridge MA 02139, USA,
nikovski@merl.com
[2] Department of Electrical and Computer Engineering , University of Florida, Gainesville, FL 32611, USA,
gr0@cise.ufl.edu

Abstract. The paper describes a method for predicting climatic time series that consist of significant annual and diurnal seasonal components and a short-term stochastic component. A memory-based method for modeling of the non-linear seasonal components is proposed that allows the application of simpler linear models for predicting short-term deviations from seasonal averages. The proposed method results in significant reduction of prediction error when predicting time series of ambient air temperature from multiple locations. Moreover, combining the statistical predictor with meteorological forecasts using linear regression or Kalman filtering further reduces prediction error to typically between 1^oC over a prediction horizon of one hour and 2.5^oC over 24 hours.

1 Introduction

Many processes of practical interest in everyday life, such as climate variation (air temperature and humidity) and electrical power demand have very significant seasonal components that are driven by natural phenomena such as the Earth's rotation around its axis and the Sun. These seasonal components render the time series non-stationary, and complicate the estimation of suitable prediction models for several major applications such as planning the generation of electricity and determination of its price, as well as the optimal scheduling of the operation of air conditioners, heating devices, domestic appliances, etc. The accurate prediction of such time series would result in efficient utilization of capital equipment, as well as positive environmental impact.

Due to the high practical significance of this class of problems, many forecasting approaches have been tried. Within the classical time series prediction methodology that is based on auto-regressive moving average (ARMA) models, a possible method for handling non-stationarity is to difference the time series as many times as necessary to make the resulting time series stationary [4]. Such models are also known as integrated ARMA (ARIMA) models. However, if the seasonal component itself is non-linear, after differencing, the resulting time series might exhibit non-linear dependencies, which would preclude the use of low-order linear prediction models for modeling.

It has also been discovered that direct application of more advanced machine learning techniques, such as neural networks, to the prediction of such time series can often result in poor accuracy, despite their high flexibility and ability to model dynamic systems [5]. This has been attributed to both numerical optimization difficulties, as well as to possible mismatches between the model and the physical process that generated the time series.

A much more physically realistic approach consists of decomposing a seasonal time series as a sum of explicit seasonal components and a random noise component, and modeling these components separately. The two models need not be of the same type: for example, the models of the seasonal component can be non-linear, while the model for the random deviations can be linear.

This approach corresponds well to the physical nature of some of the phenomena listed above. Using the classical decomposition model, we represent a time series X_t produced by one of these phenomena by a sum of a seasonal component s_t and a random noise component Y_t, in the absence of a trend [4]: $X_t = s_t + Y_t$. We then hypothesize that for the listed phenomena, the random noise component Y_t is stationary, and can be predicted solely from its past values $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-w}$ for some width w of a window of past values, and that the seasonal component has a fixed period h: $s_t = s_{t+kh}$ for all integers k. The problem then reduces to modeling well the two parts of the decomposition, s_t and Y_t.

In the remainder of the paper, we propose a memory based method for the estimation of the seasonal component s_t for the case of seasonal time series with annual and diurnal components, and present experiments on a number of data sets for ambient air temperature in multiple parts of the USA. We also describe how the predictions of the proposed model can be combined with meteorological forecasts in real time.

2 Memory-Based Modeling of Seasonality

The motivating application for this method is the prediction of ambient temperature at a specific location, for example outside of a residential or commercial building, from a database of historical readings at that location. This temperature will determine the actual thermal load that would be experienced by heating and air conditioning equipment, and is essential for the optimal scheduling of its operation. Typical prediction horizons of interest are around 24 hours.

In most areas of the world, the ambient temperature is subject to very large variations due to two cyclical components. The first one is the change of seasons caused by the rotation of the Earth around the Sun (annual component). The second one is the change of night and day, caused by the rotation of the Earth around its axis (diurnal component). In addition to these two seasonal components, a random component exists that is caused by meteorological phenomena such as cold and warm fronts, cloud cover, wind, solar activity, etc. This component is irregular, but fairly inert and persistent — such conditions usually persist for intervals commensurate with the prediction horizon of interest.

In the classical decomposition framework, the two seasonal components would be modeled separately to produce average annual temperatures and average daily temperatures, to be subtracted from the original time series in order to deseasonalize it. For a number of reasons, this approach would not work with temperature time series.

First, the daily variation of temperatures at a particular location usually does not depend only on the time of the day, but also on the day of the year. The reason is that depending on how high the Sun is, some parts of the building would be in the shadow or not, thus strongly affecting the air temperature there. (The curve traced by the Sun along the sky at the same time every day for an entire year is called *analemma*, and its vertical variation is around 46.878^o, or twice the angular tilt of the Earth.) Because of this, the two seasonal components should be modeled together, and many prediction methods estimate the average temperature for a specified combination of date and time of the day. One simple way to achieve this is to do *calendar averaging*: for any combination of date and time of the day, for example 3pm on January 23, compute the average of all readings from a historical database of temperatures that have been recorded at 3pm on January 23 of any year.

The second reason classical decomposition, including calendar averaging, would not work well, is that the period of rotation of the Earth around the Sun is not an integer number of days. Rather, the exact period of rotation is 365.25636042 solar days, also known as a *sidereal* year, i.e., measured with respect to the background stars. As it is well known, the fractional part of one quarter of a day is corrected by means of a leap year every four years.

The practical consequence of this is that the concept of average temperature at a specified combination of date and time of the day does not actually make sense. It is not correct to speak of the average temperature at 3pm on January 23, because depending on which year this day is in, the Earth might be at significantly different positions along its orbit around the Sun, and hence the impact of the Sun on the climate would be different. For example, on January 23, 2009, the Earth's position with respect to the Sun will be closer to that of January 24, 2008, rather than to that of January 23, 2008, due to the fact that 2008 was a leap year. (If February 29, 2008 did not exist, January 23, 2009 would have been dated January 24.) However, it will also be closer to that of January 23, 2005, rather than January 24, 2005.

In order to account for this mechanism, we propose an alternative memory-based estimation method called *sidereal averaging*. This method never computes explicit temperature estimates for a general combination of date and time of day; rather, it consults the database of historical readings only after a query time is given on a specific day of a specific year. The algorithm then retrieves and averages, for each year of data in the database, the temperature on the day when the position of the Earth along its orbit around the Sun was closest to its position on the query day, time, and year. In its characteristics, this algorithm is similar to other memory-based machine learning algorithms, such as k-nearest neighbors. The novelty in this algorithm is the distance measure used, that is,

the distance between corresponding positions of the Earth along its orbit around the sun.

We investigated the effect of the sidereal vs. calendar averaging methods in an experimental study, as described in Section 4. In both cases, we modeled the random component after deseasonalizing by means of low-order ARMA models.

3 Combining Statistical and Meteorological Forecasts

The methods described in the previous two sections, including the proposed method for sidereal averaging, are statistical machine learning methods: they use a database of past examples to build a predictive model, using various machine learning tools. However, for the case of temperature prediction, there is another very important source of forecasts: the governmental meteorological agencies in practically every country of the world. In recent years, detailed forecasts have been made available in real time using convenient information and communication infrastructure. For example, the National Weather Service (NWS) of the United States has been offering weather forecasts for the entire territory of the country as a standard web service since 2004. Using these forecasts in real-time prediction would be very desirable.

However, these forecasts have a significant disadvantage: they are produced for a relatively small number of locations, typically airports, and even the closest location to the target place for prediction might have significantly different weather patterns. The question, then, is how to combine the local statistical prediction with the regional meteorological forecast.

This problem has also been subject to intensive research. Kawashima et al. proposed a curve fitting method based on the high and low temperature of the forecast [2]. Shaheen and Ahmed extended the method to include the current temperature as well [3]. Linear regression methods have also been tried, for example using the form:

$$T_t = a\bar{X}_t + bZ_t + c, \tag{1}$$

where T_t is the combined forecast at time t, $\bar{X}_t$ is the temperature predicted by the statistical method, and Z_t is the temperature according to the meteorological forecast, possibly for a fairly different location. The regression coefficients a, b, and c can be estimated from a relatively small dataset of past values for the three variables T, $\bar{X}$, and Z. They can also be continuously re-estimated, for example from the values immediately preceding the current moment in time. Another variation includes regression parameters a_l, b_l, c_l that are dependent on the prediction horizon $l = 1, L$, to account for the varying ratio between the prediction errors of the statistical and meteorological forecasts that is typically encountered in practice. Here L is the longest prediction horizon, measured in time steps.

Another method for combining forecasts is based on a Kalman filter [1]. The idea is to treat the meteorological forecast as a correction factor for the local forecast, through a gain matrix K_t that is re-estimated continuously:

$$\mathbf{T}_t = \bar{\mathbf{X}}_t + K_t(\mathbf{Z}_t - H\bar{\mathbf{X}}), \tag{2}$$

where H is a selection matrix, and the variables of interest are vectors of dimensionality L: $\mathbf{T}_t = [T_{t+1}, T_{t+2}, \ldots, T_{t+L}]^T$, $\bar{\mathbf{X}}_t = [\bar{X}_{t+1}, \bar{X}_{t+2}, \ldots, \bar{X}_{t+L}]^T$, $\mathbf{Z}_t = [Z_{t+1}, Z_{t+2}, \ldots, Z_{t+L}]^T$. It can also be shown that the Kalman filter is a special case of a linear regression method, where the regression coefficients are estimated differently, and also the statistical dependency between prediction at different horizons can be modeled, too.

4 Experimental Verification of Prediction Methods

In order to train and evaluate the described prediction methods, hourly temperature data over 13 years (1995-2008) was obtained from the National Climatic Data Center (NCDC) of the USA for the eastern half of the country, for a total of 26GB of data. The NCDC data had been recorded at most international airports and some local airports in the USA, and the National Weather Service also provides online forecasts for most of the same locations in XML format. The forecasts are available for the times of the day of 02:00, 05:00, 08:00, 11:00, 14:00, 17:00, 20:00, and 23:00 hours with prediction horizon of the next 5 days. Updated observations are available hourly.

Since historical data and forecasts were available only at airports, we performed the analysis on pairs of airports, treating one of the airports as the prediction target location, and the other one as the source of meteorological forecast (and vice versa). The pairs of airports had a distance of between 30 and 50 miles between each other, in order to represent the typical distance from a residential or commercial building to the nearest airport. After examination of the data, four pairs of locations were selected for experimental verification, combining northern/southern latitude and coastal/continental climates. The pairs are shown in Table 1. The forecasts and the observed temperatures were recorded into a SQL server over a period of four weeks.

Table 1. Airport pairs for experimental verification

	Northern	Southern
Coastal	BOS Boston Logan Airport	MIA Miami International Airport
	OWD Norwood Memorial Airport	TMB Kendall Tamiami Airport
Continental	CLE Cleveland Hopkins Airport	ATL Atlanta Hartsfield Airport
	BKL Burke Lakefront Airport	FTY Fulton County Airport

After modeling and subtracting the seasonal component using either calendar or sidereal averaging, we interpolated the resulting deviations from the seasonal average at a time interval of Δt equal to 1 hour, and modeled the resulting time series of deviations Y_t by means of ARMA models. We experimented with models

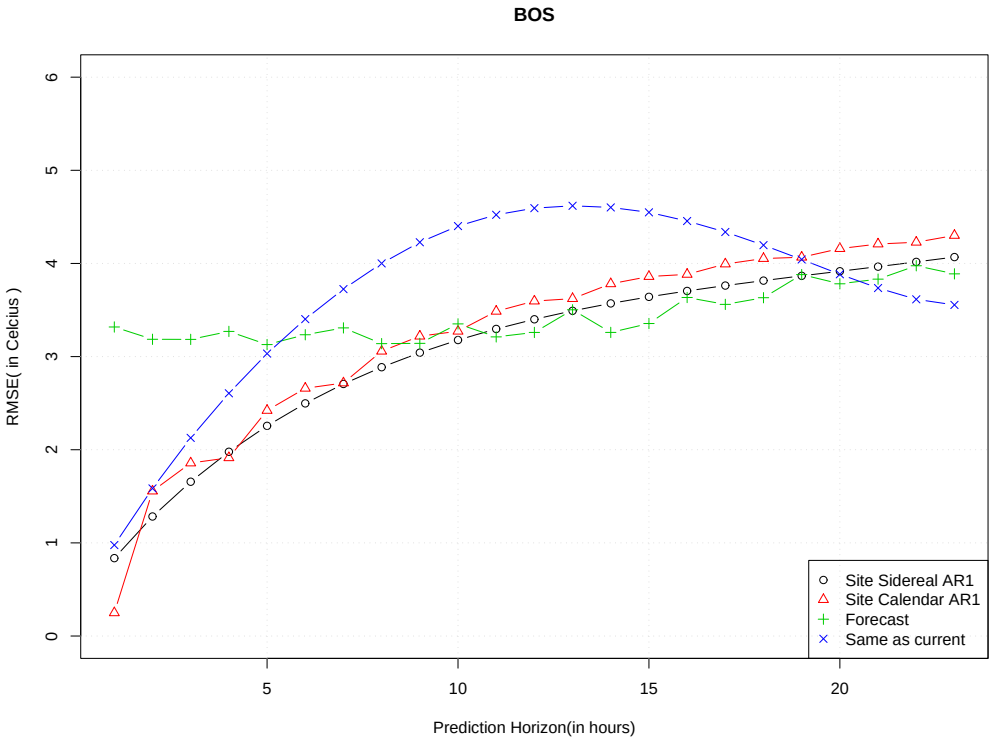
of order varying from (AR=1,I=0,MA=0) to (AR=3,I=1,MA=1) for fitting the deviations, using the Time Series package TS in the statistical environment R, and discovered that even the simplest autoregressive model AR(1) of order one was very successful at modeling the deviations. For example, its prediction error for the deviations at Boston Logan Airport after sidereal averaging was only 1.5% higher than that of the best ARIMA model. Furthermore, for the AR(1) model, the single regression coefficient r was typically very large for all time series Y_t, around $r = 0.98$. This suggests that the deviations from normal seasonal average temperatures typically persist for a fairly long time, and prediction over horizons of up to 24 hours is indeed practically possible.

As regards the relative performance of sidereal vs. calendar averaging, Figures 1 through 4 show that in all cases sidereal averaging is either much better than calendar averaging (OWD, TMB, CLE, ATL, BKL), or the same (BOS, MIA).

Figures 1 through 4 also show an interesting pattern — although the prediction error of statistical prediction methods increases with the time horizon, as expected, the error of the meteorological forecast does not. This can be explained by the completely different methodology used by meteorological agencies, but still the accuracy of such forecasts at relatively long prediction horizons is remarkable. At the same time, their accuracy at short prediction horizons is much worse than that of the statistical predictors, which is an excellent justification for methods that attempt to combine statistical and meteorological forecasts.

Regarding the improvements in accuracy that can be achieved by such methods, Figures 5 through 8 show that in most cases (BIOS, CLE, ATL, BKL) the combined forecast by either linear regression or a Kalman filter is much more accurate than either one of the individual forecasts. In some cases (FTY, TMB, MIA) there is no significant difference for horizons longer than 6 hours, but still the combined prediction is more accurate for horizons shorter than that, and there is only one single site (OWD) where the combined predictions are significantly less accurate than the meteorological forecast. Furthermore, systematic and significant differences between the performance of linear regression and Kalman filtering cannot be observed, which means that linear regression should be preferred in practical systems for its ease of implementation. (In all of these experiments, only sidereal averaging was used for combination with the meteorological forecast, since the previous set of experimental results established its superior performance over calendar averaging.)

In absolute terms, the performance of the combined predictors can be summarized as follows. Accuracy starts around 1^oC to 1.5^oC for a prediction horizon of 1 hour, and grows to between 2^oC and 2.5^oC during the next 3 to 10 hours, almost never exceeding 3^oC over the longest prediction horizon of interest, 24 hours. Compared with the variability of daily and annual temperatures which span an interval of around 50^oC for most climates, this accuracy can be considered practically very useful.



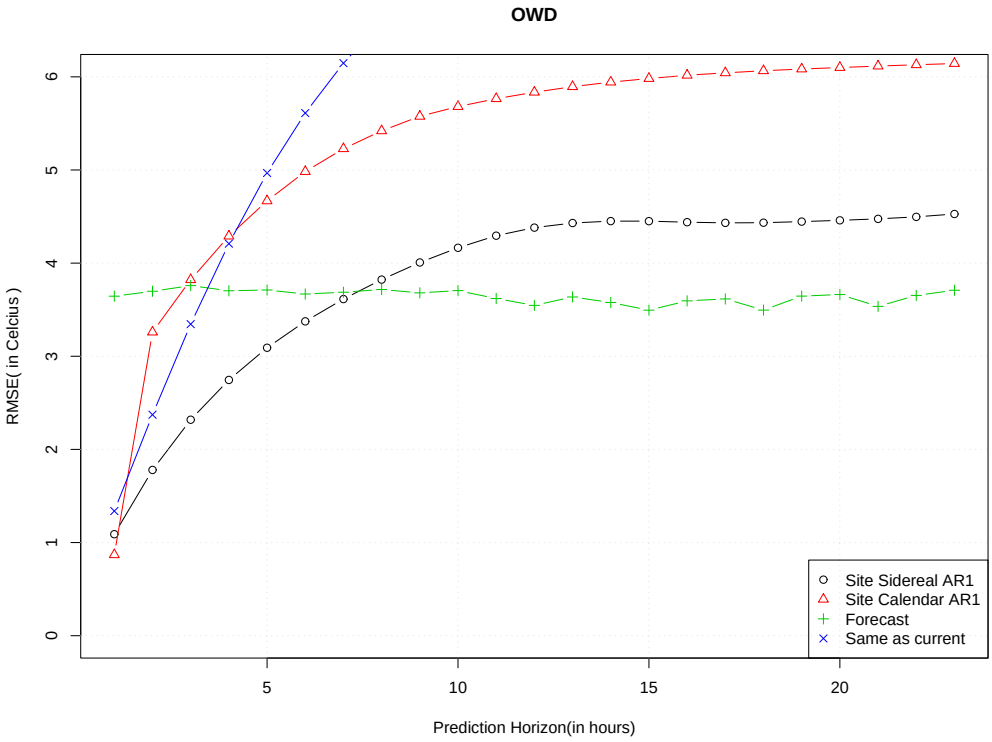
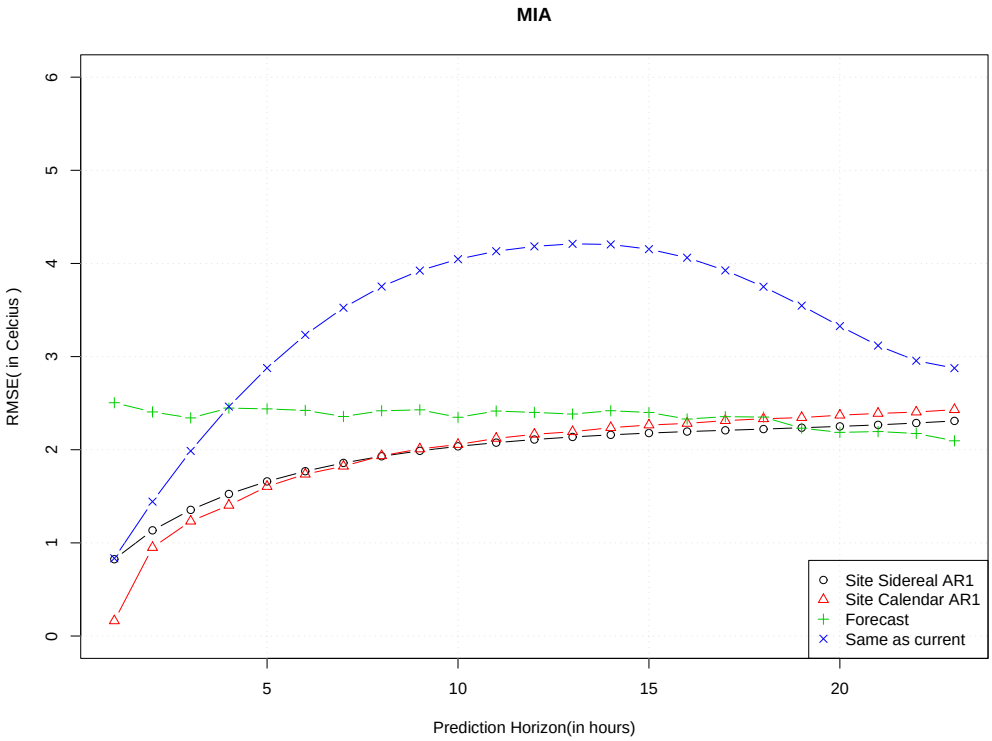


Fig. 1. Comparison between seasonal averaging methods in a northern coastal climate.



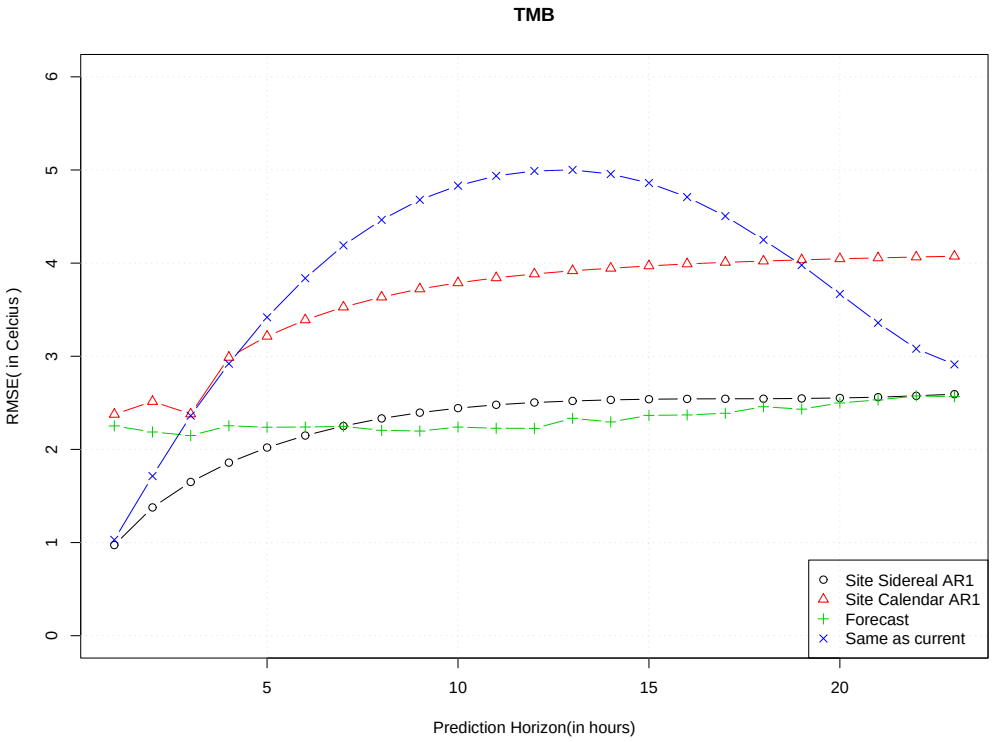
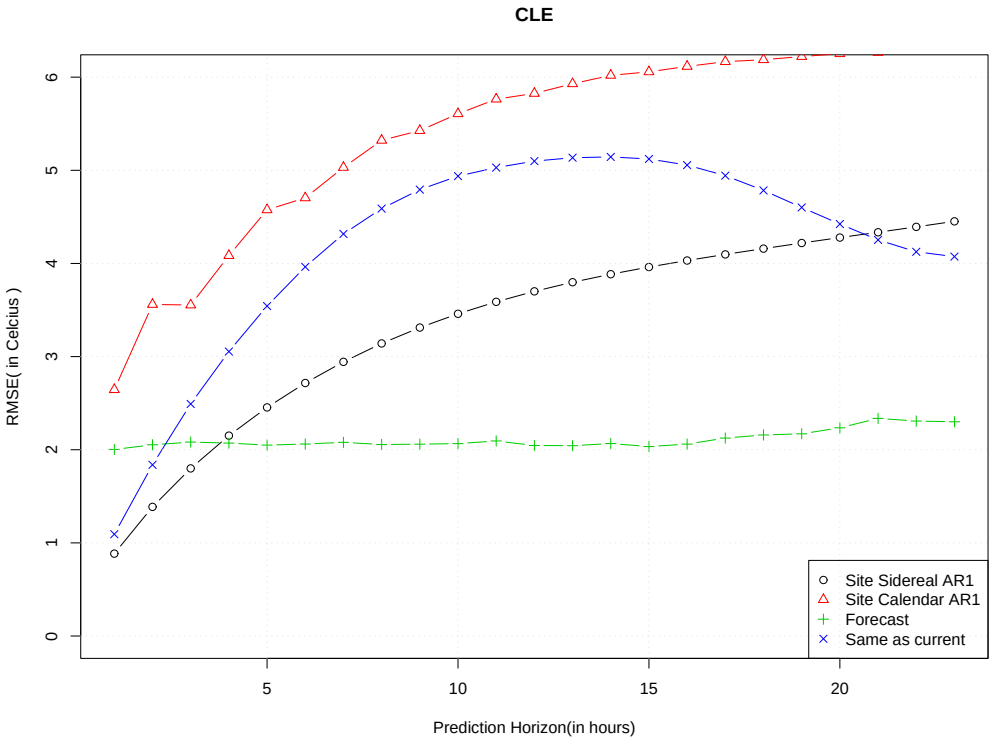


Fig. 2. Comparison between seasonal averaging methods in a southern coastal climate.



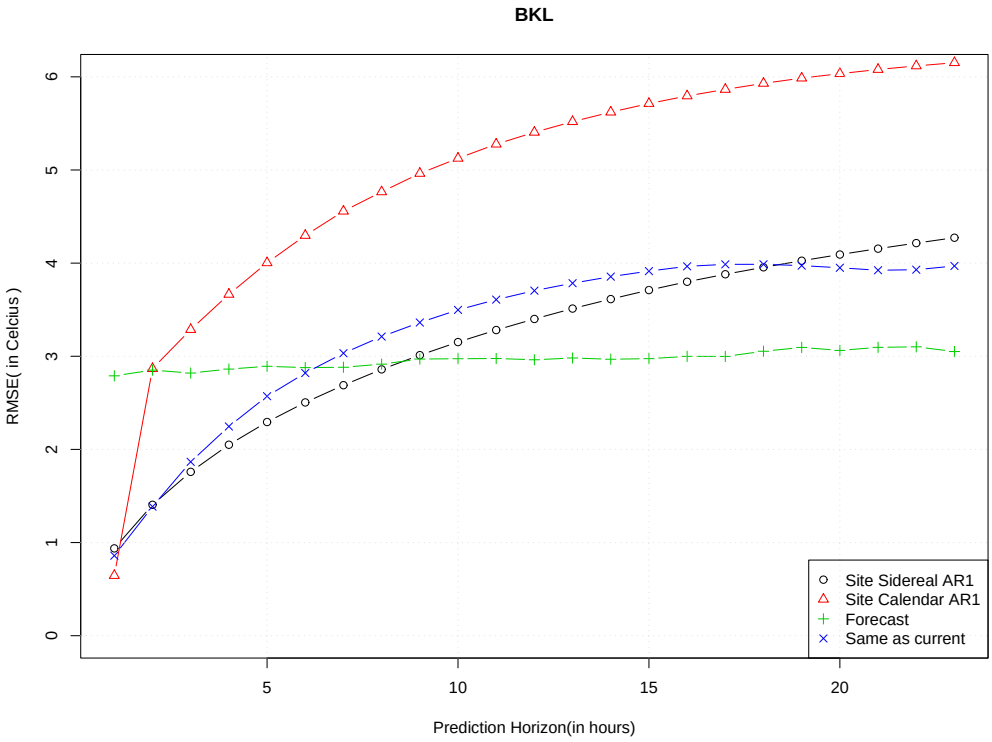
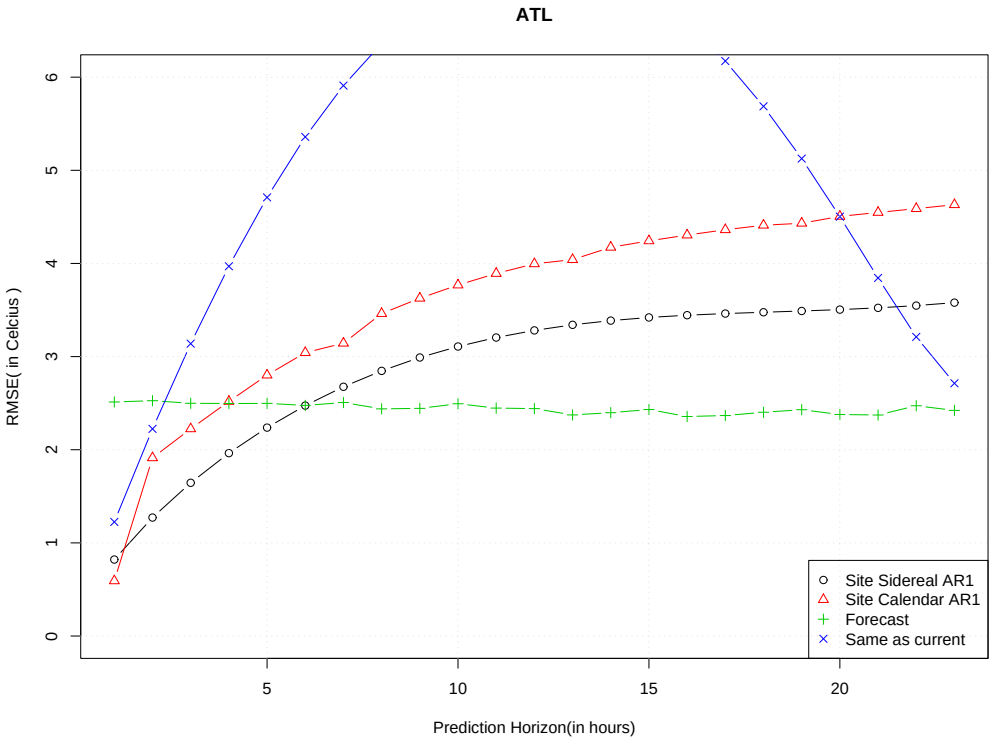


Fig. 3. Comparison between seasonal averaging methods in a northern continental climate.



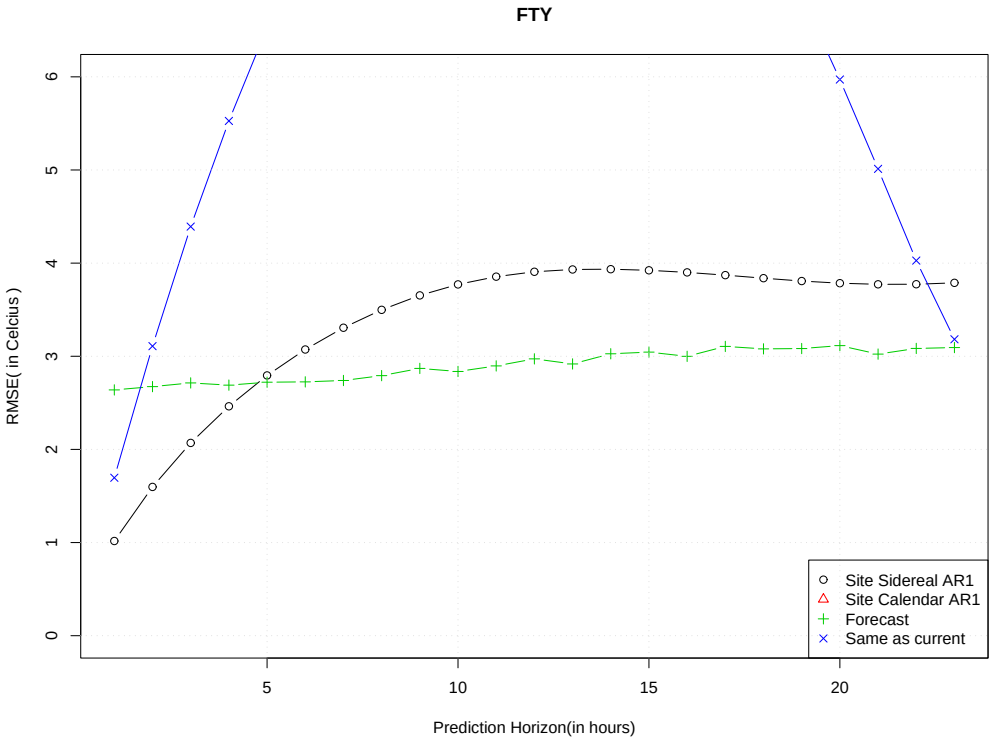
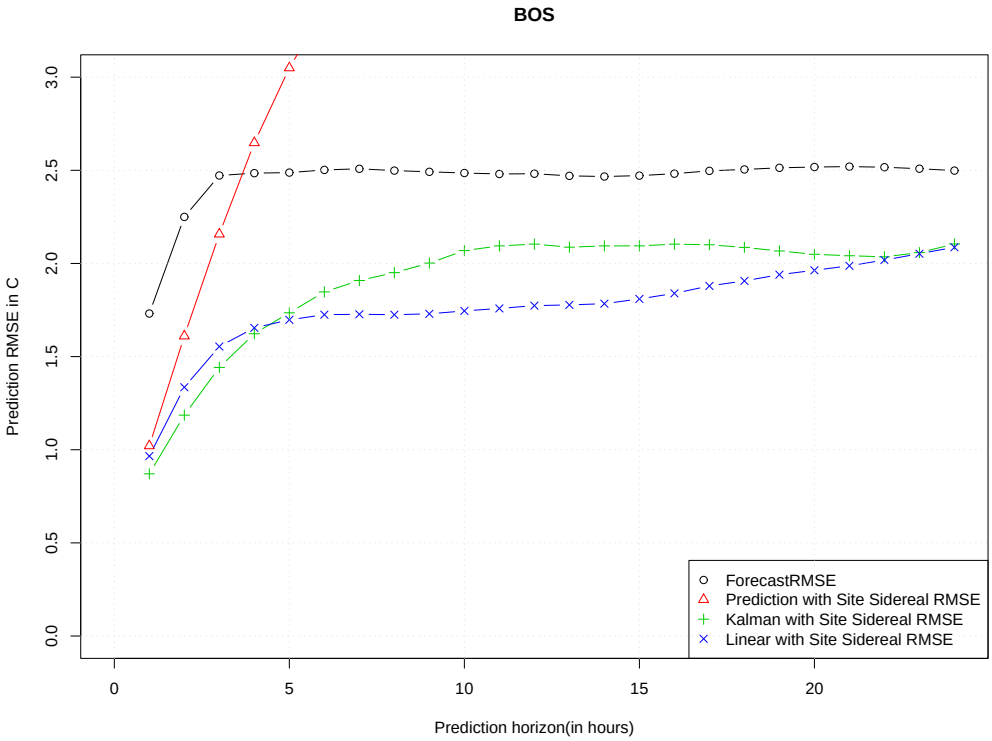


Fig. 4. Comparison between seasonal averaging methods in a southern continental climate.



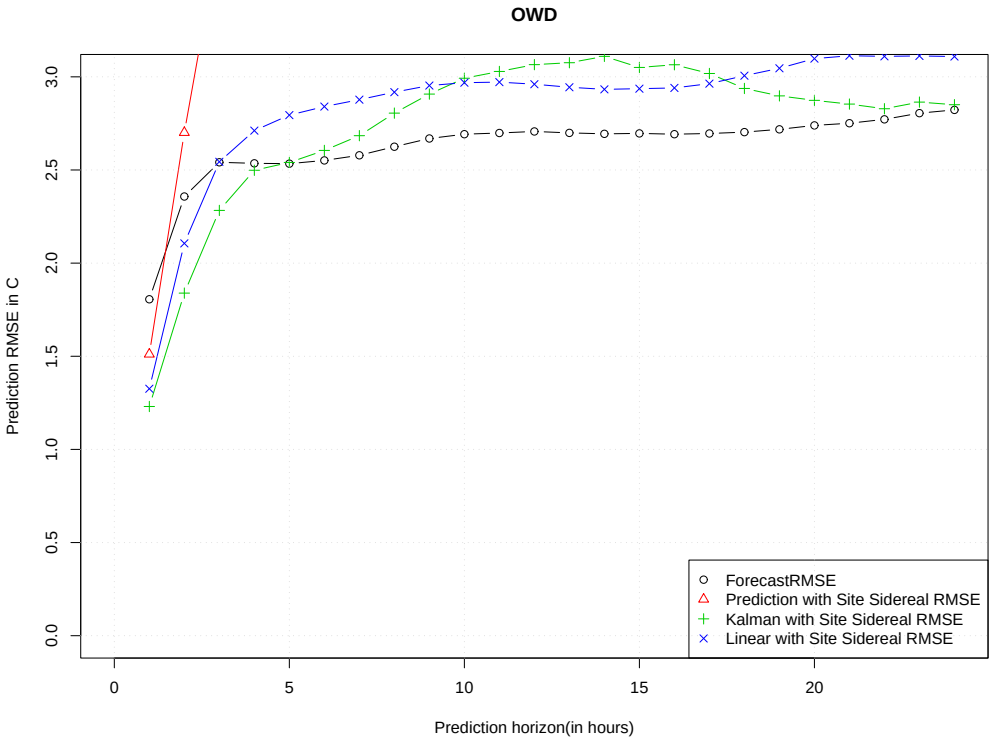
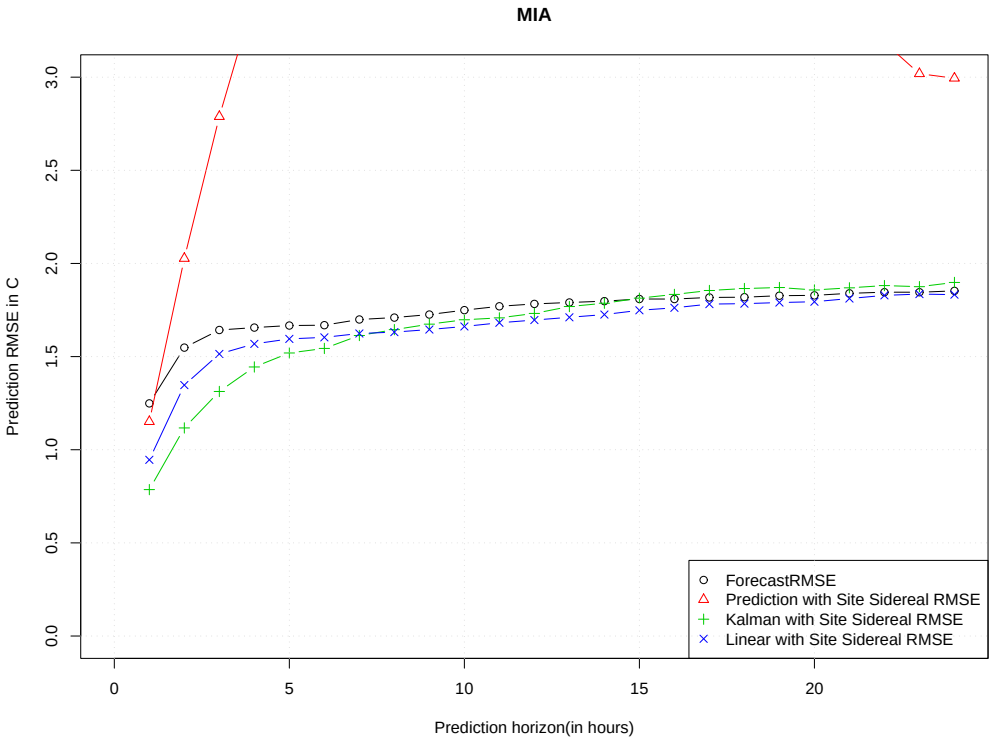


Fig. 5. Comparison between forecast combination methods in a northern coastal climate.



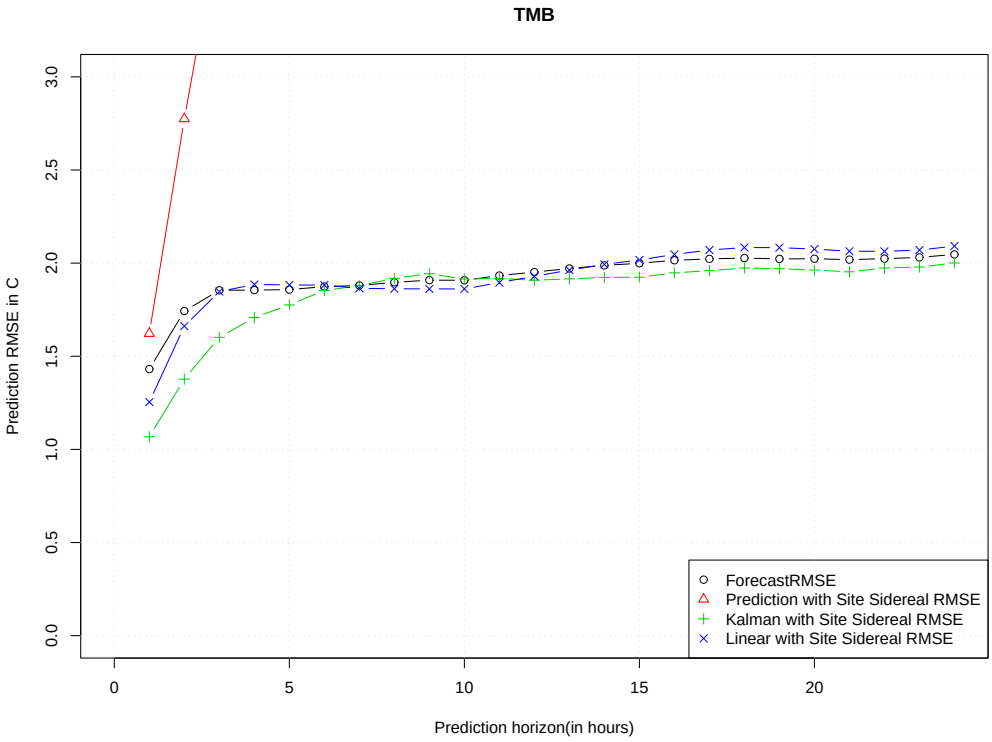
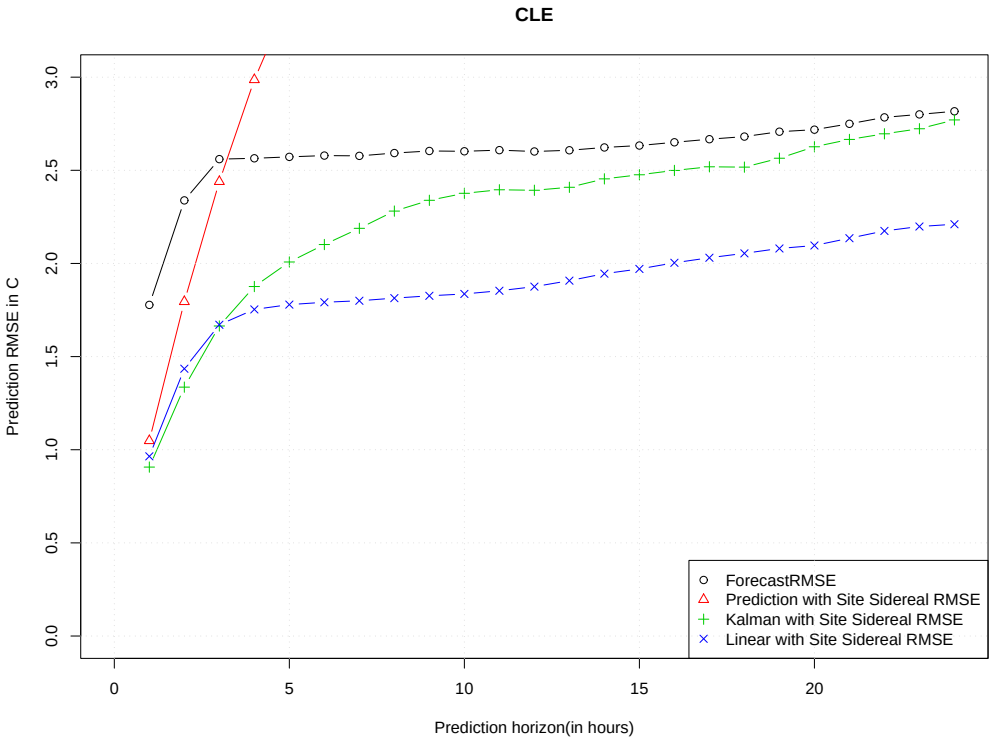


Fig. 6. Comparison between forecast combination methods in a southern coastal climate.



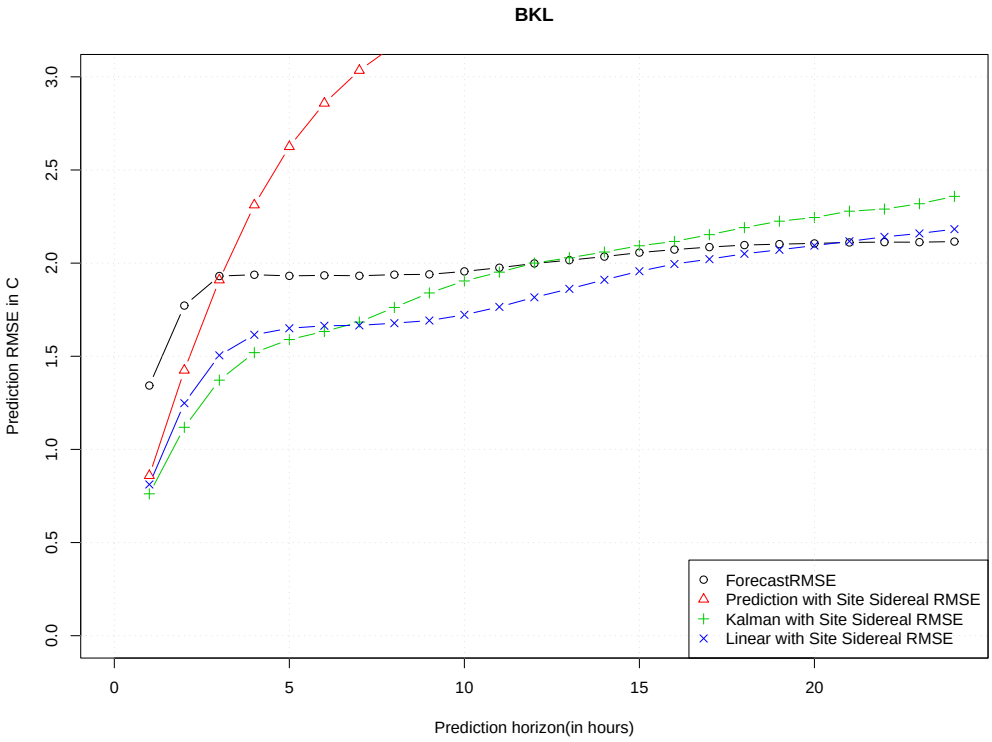
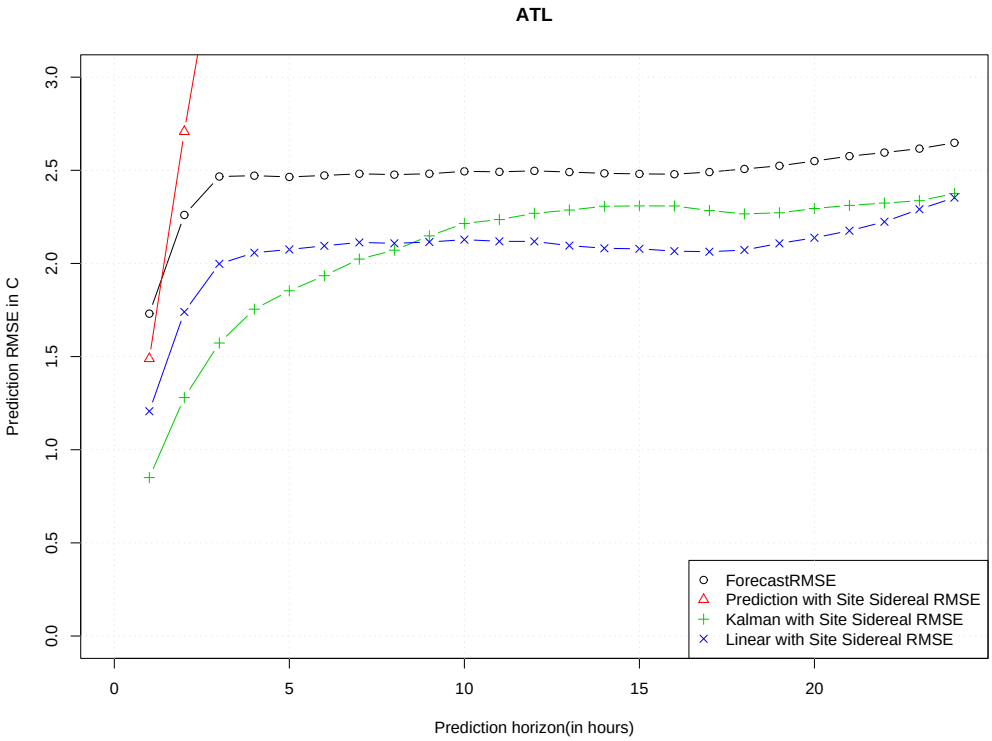


Fig. 7. Comparison between forecast combination methods in a northern continental climate.



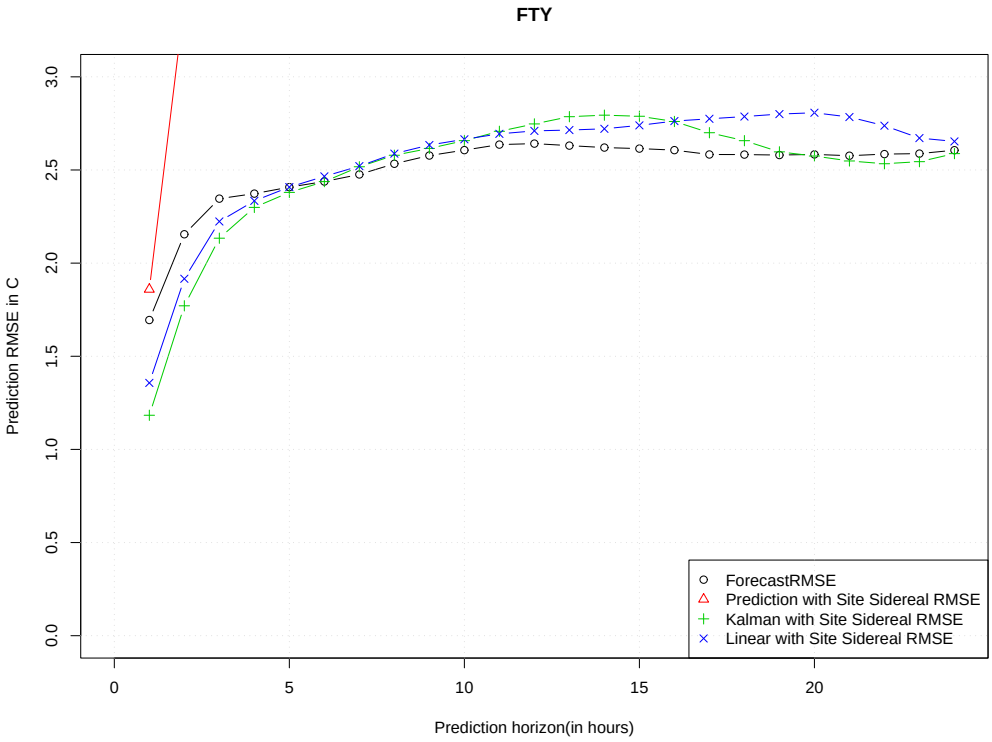


Fig. 8. Comparison between forecast combination methods in a southern continental climate.

5 Conclusion and Future Work

We have proposed a novel method for modeling of the annual and diurnal seasonality of time series, and have demonstrated experimentally that it significantly improves the accuracy of prediction of the temperature of ambient air when combined with low-order ARMA models of the deseasonalized time series. The method is similar to other memory-based machine learning techniques such as k-nearest neighbors (kNN), and uses as distance function the difference between pairs of positions of the Earth's along its orbit around the Sun. (In this case, the number of neighbors is equal to the number of years for which training examples have been collected.) Due to this similarity, it might be expected that algorithms that are counterparts to other memory-based machine learning methods, such as locally weighted polynomial regression, or kNN with more neighbors might improve the accuracy of seasonal modeling even further [6]. In practice, this would mean including observations from more days into the averaging process, possibly using variable weights. Also note that this experimental analysis does not prove that the sidereal averaging method is better than the calendar method for the purposes of modeling of seasonal components; it merely indicates that the sidereal method is significantly better when followed by low-order ARMA modeling of the remaining random component.

We have also demonstrated that two linear methods for combining of the statistical prediction with meteorological forecasts further reduces prediction error significantly. In these experiments, the meteorological forecast from the nearest airport was always used, but this approach can be extended in the future to using forecasts from multiple locations, which might improve accuracy even further in densely populated areas.

Finally, the described method should also be applicable to other time series whose dynamics are determined entirely or partly by the motion of the Earth and the Sun, for example daily light, humidity, electrical power demand, etc.

References

1. Nagai, T., A method for revising temperature and humidity prediction using additional observations and weather forecasts, Proceedings of Building Simulation 2007, pages 245-252.
2. Kawashima, M, Dorgan, C.E, and Mitchell, J.W., Hourly thermal load prediction for the next 24 hours by ARIMA, EWMA, LR, and an artificial neural network, ASHRAE Transactions, Vol.101(1):186-200, 1995.
3. Shaheen, N.I. and Ahmed, O., A simple methodology to predict local temperature and humidity, ASHRAE Transactions, Vol.104, Part 1A, pp.451-459, 1998.
4. Brockwell, P. and Davis, R. Introduction to Time Series and Forecasting. Springer, second edition, 2002.
5. Faraway, J. and Chatfield, C. Time series forecasting with neural networks: a comparative study using the air line data. Journal of the Royal Statistical Society: Series C (Applied Statistics), Vol. 47, No. 2, pp. 231-250
6. Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning, Springer, 2001.