University of Alberta

Modeling Uncertainty of Numerical Weather Predictions Using Learning Methods

by

Ashkan Zarnani

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems
Department of Electrical and Computer Engineering

Spring 2014
Edmonton, Alberta

To

The memory of my dad…

Abstract

Weather forecasting is one of the most vital tasks in many applications ranging from severe weather hazard systems to energy production. Numerical weather prediction (NWP) systems are commonly used state-of-the-art atmospheric models that provide point forecasts as deterministic predictions arranged on a three-dimensional grid. However, there is always some level of error and uncertainty in the forecasts due to inaccuracies of initial conditions, the chaotic nature of weather, etc. Such uncertainty information is crucial in decision making and optimization processes involved in many applications. A common representation of forecast uncertainty is a Prediction Interval (PI) that determines a minima, maxima and confidence level for each forecast, e.g. [2℃, 15℃]-95%.

In this study, we investigate various methods that can model the uncertainty of NWP forecasts and provide PIs for the forecasts accordingly. In particular, we are interested in analyzing the historical performance of the NWP system as a valuable source for uncertainty modeling. Three different classes of methods are developed and applied for this problem. First, various clustering algorithms (including fuzzy c-means) are employed in concert with fitting appropriate probability distributions to obtain statistical models that can dynamically provide PIs depending on the forecast context. Second, a range of quantile regression methods (including kernel quantile regression) are studied that can directly model the PI boundaries as a function of influential features. In the third class, we focus on various time series modeling approaches including heteroscedasticity modeling methods that can provide forecasts of conditional mean and conditional variance of the target for any forecast horizon.

All presented PI computation methods are empirically evaluated using a developed comprehensive verification framework in a set of experiments involving real-world data sets of NWP forecasts and observations. A key component is proposed in the evaluation process that would lead to a considerably more reliable judgment. Results show that PIs obtained by the ARIMA-GARCH model (for up to 6-hour-ahead forecasts) and Spline Quantile Regression (for longer leads) provide interval forecasts with satisfactory reliability and significantly better skill. This can lead to improvements in forecast value for many systems that rely on the NWP forecasts.

Acknowledgements

Firstly, I would like to thank my supervisor Dr. Petr Musilek for granting me the incredible opportunity to conduct this exciting research with his support and guidance. Secondly, I would like to thank FACIA members and other peers in the University of Alberta for their input, collaboration and friendship. I would also like to thank my dear friends here in Edmonton, Pooya and Mona, Reza and Mana, Meysam and Fatemeh and other friends as well that made my life as a graduate student much more fun and supported me through these years. Finally, I would like to thank my dearest of all, my wife, Hamideh who has been the greatest partner and whose love has always been the defining factor and an endless motivation for my life. I can never put into words how much I appreciate all the efforts and support from my parents and my brothers. Particularly, my brother Saman, who has always been most supportive, graciously sacrificing and a perfect role model for me.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACF	Auto Correlation Function
AG	Agassiz-Hope Data Set
AR	Autoregressive
ARCH	Autoregressive Conditional Heteroscedasticity
ARIMA	Autoregressive Integrated Moving Average
BC	British Columbia – Data Set
BIC	Bayesian Information Criterion
CDF	Cumulative Distribution Function
DTR	Dynamic Thermal Rating
EACF	Extended Auto Correlation Function
FCM	Fuzzy C-means
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
HClust	Hierarchical Clustering
KQR	Kernel Quantile Regression
LocQR	Local Quantile Regression
LQR	Linear Quantile Regression
LSE	Least Squares Estimation
MA	Moving Average
MA-D3	Baseline model: Moving average using past 3 days
MA-H4	Baseline model: Moving average using past 4 hours
MLE	Maximum Likelihood Estimation
NLQR	Non-Linear Quantile Regression
NWP	Numerical Weather Prediction
PACF	Partial Auto Correlation Function
PCA	Principal Component Analysis
PDF	Probability Density Function
PI	Prediction Interval
QR	Quantile Regression
RMSE	Root Mean Squared Error
sARIMA	Seasonal ARIMA
SPQR	Spline Quantile Regression
SScore	Skill Score
sxARIMA	Seasonal Exogenous ARIMA
WRF	Weather Research and Forecasting

List of Notations

$\langle .,. \rangle$	Inner product of two vectors
β_y^θ	Vector of quantile regression coefficients for the θ-quantile of target *y*
B	Number of bootstrap samples
C	Cost factor of KQR objective
$\bar{\delta}_M^\alpha$	Average distance of missed observation from PI boundaries forecasted by method *M*
$\bar{\Delta}_M^\alpha$	Average distance of observations from PI forecasts of method *M*
$\bar{\Delta}_M^{\alpha,j}$	True average distance of observations from PI forecasts of method *M* in cluster *j*
$\widehat{\bar{\Delta}_M^{\alpha,j}}$	Estimated average distance of observations from PI forecasts of method *M* in cluster *j* in test experiments
c_i	Cluster representative (centroid) for cluster *i*
d	Number of forecast attributes (influential variables)
df	Degrees of freedom in the B-spline basis function
D	Set of all available forecast cases (vectors) in train set
D_i	Set of forecast cases (vectors) in cluster *i*
E_i^y	Set of forecast errors for cases in cluster *i*
e_j^y	Error of forecast case *j* for target *y*
e_t	Forecast error at time *t*
ϕ_i	Coefficients of an AR process
$\phi(\mathrm{x})$	Transformation (mapping) function of the x vector
$f_{j,k}$	B-spline basis function for feature *j* using *k* degrees of freedom
$\hat{f}_{y_t}$	pdf of a forecast target y_t
$\hat{F}_t^e$	Estimated cdf of forecast error
$\hat{F}_{y_t}$	cdf of a forecast target y_t
$\gamma_{t,s}$	Covariance between Y_t and Y_s
h	Smoothing parameter in kernel density smoothing
K	Number of clusters (as a parameter of the clustering algorithm)
K_{ij}	Element *i*,*j* of the kernel matrix *K*
$K(.)$	Kernel density function in kernel density smoothing
$K(.,.)$	Kernel function
λ	Neighbourhood parameter of LocQR
l	Forecast horizon – lead time
L_θ	Loss function of a θ-quantile
$\hat{L}_{\hat{y}_t}^\alpha$	Lower bound of the prediction interval over forecast target y_t with confidence level of 1- α

μ_t	Mean of the stochastic process *Y* at time *t*
$\hat{\mu}_e$	Sample mean of forecast error
$\hat{\mu}_e^i$	Sample mean of forecast error in cluster *i*
m	Fuzzification factor in FCM
N	Number of training samples – available hourly point forecast errors
p	Order of an AR process
P	Order of seasonal AR terms
q	Order of an MA process
$q_{y_t}^{(\alpha)}$	α-quantile of y_t
$\hat{q}_{y_t}^{(\alpha)}$	Estimation of α-quantile of y_t
Q	Order of seasonal MA terms
ρ_k	Autocorrelation at lag *k*
$\rho_k(X,Y)$	Cross-correlation between random variables *X* and *Y*
$\rho_{t,s}$	Correlation between Y_t and Y_s
r_k	Sample autocorrelation at lag *k*
$r_k(X,Y)$	Sample cross-correlation between random variables *X* and *Y*
r_t	The time series of the residuals of another time series model
Rel_M^α	Reliability of method *M* in test experiments for PIs with confidence level of 1-α
Res_M^α	Resolution of method *M* in test experiments
σ	Gaussian kernel function parameter
$\hat{\sigma}_e$	Sample standard deviation of forecast error
$\hat{\sigma}_e^i$	Sample standard deviation of forecast error in cluster *i*
$\sigma_{t\|t-l}^2$	Conditional variance forecast for time *t* by forecast horizon *l*
s	Seasonal period
Shp_M^α	Sharpness of method *M* in test experiments
$SScore_M$	Skill score of method *M* in test experiments
$SScore_M^{0.95}$	95% bound on skill score of method *M* in test experiments
Ψ_j	Coefficients of a General Linear Process
θ_i	Coefficients of an MA process
t	Time
T_j	Number of test samples in cluster *j*
$t(\alpha,n)$	The quantile of the Student's t-distribution for confidence level α and n degrees of freedom
u_{ij}	Degree of membership of the point x_i in cluster *j*
$\widehat{U}_{\hat{y}_t}^\alpha$	Higher bound of the prediction interval over forecast target y_t with confidence level of 1- α
ξ_i^I	Hit indicator for prediction interval *I*
x_j	Vector number *j* of forecast attributes
x_{new}	Vector of forecast attributes for a *new* forecast at test time
y_t	Observation of target y_t
$\hat{y}_t$	Point forecast (expected value) of target y_t

Y_t	Time series value at time *t* of the stochastic process *Y*
$\hat{Y}_t(l)$	Forecast for Y_{t+l} made at time *t*
z_α	Two-tailed critical Z-Value for normal distribution with significance level of α

Chapter 1
Introduction

This chapter lays a starting point into the study conducted in this thesis.

1.1. Forecasting and Uncertainty

Weather prediction has numerous applications in various domains. Weather forecasts are typically made and reported in the form of an expected value for the attribute of interest in a particular time and location. Numerical weather prediction (NWP) models are advanced computer simulation systems that provide such expected value forecasts for a number of attributes. They capture physical atmospheric processes to model the atmospheric behavior. Although the deterministic interactions of these physical simulations yield real numbers (with even third decimal places) of different weather attributes in the mid-range future, such forecasts are uncertain due to the inaccuracy of initial conditions, low spatial resolution, and various simplifying assumptions [48][61][62]. Yet, such uncertainty information is not available in the immediate outputs of the system.

Thus, while the NWP deterministic outputs are real numbers, they will have an unavoidable level of uncertainty. As an example an NWP model may predict the temperature to be -3.8°C for a specific location both in the next hour and the next week. However, the uncertainty level of these forecasts would be clearly different although the single value (point) predictions are equal.

In many applications, it is desirable that forecasts be accompanied by the corresponding uncertainties. Information about forecast uncertainty may be as significant as the forecast itself. Such information can have important role in the planning and decision making processes that utilize the forecasts [16][72]. For instance, the expected accuracy of NWP wind speed and temperature forecasts can have crucial impact on the

optimized operational planning and management of power grids using Dynamic Thermal Rating (DTR) systems [42].

The uncertainty of a forecast is typically formulated and communicated using *prediction intervals* (PIs) that are accompanied by a percentage expressing the level of confidence (nominal coverage rate) (e.g., T = [2°C, 14°C] conf = 95%) [16][29] or simply expressed as $-3.8^{\pm 6}$, when as typical the units and confidence levels are assumed invariant for all PI forecasts of a particular system and application. The confidence level specifies the expected probability of the actual observation to be inside the PI range. This form of forecast (sometimes called a *central credible interval forecast* or *forecast interval*) may be harder for a non-specialist to interpret and evaluate, but it provides the user with a more complete description of the predicted phenomenon compared to a point forecast. Chatfield [17] classifies forecasts of real valued variables into point forecasts, interval forecasts and full density forecasts. In contrast to point forecasts, interval forecasts supply the likely uncertainty in the prediction and are therefore preferred. In spite of the clear value of PI forecasts, this format of forecast "…has been largely overlooked by meteorologists and would benefit from some attention…" [39].

When the PIs of a forecasting system take different widths depending on the forecast context dynamically [17] [68], they are called Conditional PIs as opposed to a static interval forecasting system that has the same width in any occasion. Assuming a fixed confidence level for the output PIs of a system, a forecast case with a lower level of uncertainty would in effect be narrower compared to a more uncertain forecast (e.g. $-3.8^{\pm 3}$ vs. $-3.8^{\pm 6}$).

In a deterministic forecasting system, PIs can be achieved by theoretically formulating the error behavior of the model. Such approaches are mostly infeasible due to the high degree of complexity in input data, model elements and non-linear relations between the different factors [16][20]. In many cases such error formulations turn out to be too rough and misleading [16][66].

A major category of solutions for uncertainty analysis and prediction interval estimation, especially in meteorology, is based on ensemble prediction [20]. In this method, individual predictors as members of an ensemble of forecasters, are run with different parameters and/or initial conditions [72][81]. The degree of uncertainty for a forecast is then associated with the extent of spread among these members. However, ensemble executions of an NWP model incur a very high computational cost making the

ensemble approach infeasible in many applications especially when new uncertainty analysis is required in short temporal intervals. Moreover, the ensemble approach may not be the best choice when the user is interested in the forecast uncertainty of a few points rather that a whole spatial domain.

PIs can also be obtained by statistical modeling of forecast error using the historical performance of forecasts made by the system [17][29][39][40]. In this approach, the characteristics and dynamics of the forecast uncertainty will be essentially learned from the recorded accuracy of past forecasts which are available for many deterministic forecasting systems today. In the current study, we focus on this approach as a potentially efficient method that has received relatively less attention in the literature.

The weather research and forecasting (WRF) model [67] is the NWP system used in many meteorological applications. As this model is based on deterministic formulations, it provides predictions in the form of *point* forecasts values of various weather attributes such as temperature, wind speed, wind direction, and liquid water content [48][67]. However, past forecasts made by this system (e.g. for few years) can be recorded and later augmented by actual weather observations. As it consists of forecasts and observations made at different times and locations, this historical forecast performance data can provide valuable knowledge for uncertainty analysis.

It is a well-known fact that, the extent of forecast uncertainty varies with the weather situation [62]. For example, low pressure systems are known to be less predictable than the more stable high pressure systems. It is expected that such patterns of dependency of uncertainty on the forecasted attributes can be discovered from the historical performance of the NWP forecasts [48][59][69].

In [47][48] and [49], such dependencies are discovered by clustering the performance records into separate groups and characterizing the attributes of their error distribution individually. However, this useful analysis is not practically used and evaluated for the purpose of deriving PIs from a deterministic forecasting system. In another series of studies [66][68], wind energy forecast records are grouped by an expert-driven manual partitioning of space of variables that are believed to be associated with the forecast error (influential variables). Due to the low scalability of this manual records grouping method only two variables are used to define four classes of forecast weather situation. PIs are then computed using the empirical quantiles of the error distributions in each group by using the fuzzy membership values of a new forecast in each of the predefined groups.

Experimental evaluation of the resulting PIs demonstrated applicability of the historical forecast grouping approach as it provides skilful and relatively reliable PIs from the initial point forecasts. Yet some drawbacks of this method which motivate the proposed research are discussed in the next chapter.

The application of intelligent post-processing techniques to analyze and model the uncertainty of weather forecasting systems is considered as a significant and attractive direction in probabilistic forecasts. Such methods can efficiently enrich many existing forecasting systems by providing valuable information about the uncertainty of predictions using the performance history widely accessible for these systems. In this project, we investigate the application of different learning methods to obtain effective forecast uncertainty models.

1.2. Motivation and Problem

Long records of point forecast accuracy are available in many areas specifically in NWP applications. The characteristics and dynamics of the forecast uncertainty can be essentially learned from this database of past forecasts. It is a well-known fact that, the extent of forecast uncertainty varies with the weather situation and is dependent on many factors. Such patterns of connection between forecast uncertainty and various factors can be potentially discovered in the performance history of the system.

Upon obtaining an effective model that can predict the uncertainty of NWP forecasts we can extend these forecasts as prediction intervals that can communicate such uncertainty information and provide these interval forecasts in situations when ensemble models are unavailable or infeasible. This uncertainty information is of critical significance in many decision making and optimization applications such as Dynamic Thermal Rating and energy markets just to name a few. Hence, the problem focused in this research is: “Learning and evaluation of prediction interval models of uncertainty from the historical performance of NWP forecasts.”

In particular, to improve the quality of the resulting PIs and also to alleviate the problem of manual grouping of the weather forecasts we investigate the application of automatic objective-based clustering algorithms to achieve optimally defined forecast record groups that follow the inherent structures in data. We suggest that, as these clusters will be based on the actual similarities between the past forecast situations they will lead to PIs of higher quality. Moreover, we will not be bound to the limitations of expert-based definition of partitions which becomes a daunting task with increased

dimensionality of the influential variables. In this process, we examine the application of crisp clustering algorithms i.e. K-means, CLARA and Hierarchical Clustering and assess the resulting PIs. Also Fuzzy C-means clustering is applied as a natural alternative to the crisp allocation of forecast records to a specific cluster only. The PI analysis follows a preceding step which involves the fitting of an appropriate probability distribution function to the actually observed error distribution in each cluster. We examine statistical techniques in this regard and consider the required modifications when the fuzzy approach is used.

As another alternative for learning uncertainty models, the application of Quantile Regression algorithms will be studied for the explicit and direct learning of appropriate quantile functions from the historical errors of the NWP system. In addition, time series modeling methods are investigated to account for the temporal qualities of the forecast error in the uncertainty modeling process. Inherent in all of these models, is the dynamic calibration of forecasts by modeling and removal of the "situation-based" forecast bias.

The applicability and quality of the resulting PIs in practical scenarios is investigated in this research. The results can provide insight into the role of different aspects such as clustering algorithms, number of clusters, feature sets, distribution fitting algorithms and their appropriate choice in the uncertainty modeling process. In addition, better skill and quality of the output PIs compared to some baseline PI approaches and raw point predictions of the WRF NWP system would prove the advantages and actual value of the proposed models. Some major research directions are also recognized to improve the interval forecasts by developing more appropriate learning methods.

1.3. Contributions

Major contributions of this work can be summarized as follows:

(I) We propose a new approach consisting of clustering and distribution fitting to learn models of forecast uncertainty from the NWP performance history.

(II) As a very critical aspect, we introduce a novel evaluation framework which considers sampling uncertainties in the assessment of prediction intervals in testing experiments.

(III) We develop and apply a hybrid clustering and kernel quantile regression modeling approach to the NWP prediction interval forecasting problem.

(IV) We study the application of time series modeling methods that focus both on expected conditional mean as well as the conditional variance (Heteroscedasticity) of the forecast target.

(V) The last contribution of this thesis is the broad set of experiments and analysis for the application and evaluation of a wide range of available and proposed methods for forecast uncertainty modeling using real-world large size data sets.

1.4. Organization of the Thesis

In the next chapter, basic definitions and background for the focused problem are provided. Also various approaches and previous related work is reviewed. Chapter 4 describes the various proposed clustering methods of prediction interval modeling including fuzzy C-means. To assess the quality of prediction interval forecasting models various measures and scores are incorporated into a framework described in Chapter 3. This evaluation framework is utilized in the experimental studies conducted in Chapter 4 and other chapters as well. In Chapter 5, various methods of quantile regression are studied as methods of prediction interval modeling. As a rather different approach, various time series modeling methods are investigated as an alternative solution for the problem of prediction interval modeling in Chapter 6. The experimental results and analysis of each of these groups of methods are provided within the corresponding chapter. Finally, in Chapter 7 general conclusions and future research directions are discussed.

Chapter 2
Background and Related Work

Basic definitions and related previous works are reviewed in this chapter.

2.1. Prediction Interval Definition

An NWP model or any other *deterministic* forecasting model (e.g. neural network, decision tree, etc.) provides forecasts as single values for every prediction instance. For an NWP system these forecast values are in the form of time series $\{\hat{y}_t\}(u, w, z)$ for each weather attribute (e.g., temperature, wind speed) and for each location on a three-dimensional grid (u, w, z). For simplicity, the location coordinates are omitted in the following text. However, in a *probabilistic* forecasting model the prediction is supplied as a Probability Distribution Function (pdf) $\hat{f}_{y_t}$ as an estimation of the true pdf f_{y_t} where y_t is the target variable of interest. In fact, it is expected that the actual observation y_t would be a sample following the prediction distribution $\hat{f}_{y_t}$.

A point forecast ($\hat{y}_t$) is in effect a single value from the full prediction distribution which is often selected to be the mean of this distribution as the expected value for y_t. Figure 2.1 demonstrates the distinction between a probabilistic forecast versus a deterministic forecast.

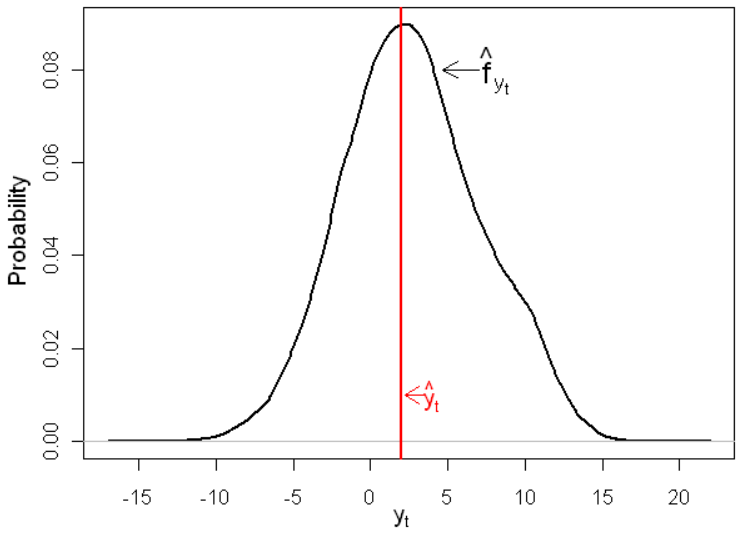


Figure 2.1. A sample probabilistic forecast reports $\hat{f}_{y_t}$ while the deterministic forecast provides $\hat{y}_t$ only

The relation between the forecast $\hat{y}_t$ and its observation y_t can be described as:

$$y_t = \hat{y}_t + e_t \qquad (2.1)$$

i.e., each observation can be decomposed to the predicted value $\hat{y}_t$ for time *t*, and an error term e_t for the specific forecast instance.

Based on a probabilistic forecast, the probability density function (pdf) $\hat{f}_{y_t}$ and the cumulative distribution function (cdf) $\hat{F}_{y_t}$ are explicitly available and are estimations of the true pdf and cdf functions of the observations f_{y_t} and F_{y_t}. Thus, an α-quantile of y_t can be defined as [29][66][68]:

$$q_{y_t}^{(\alpha)} = F_{y_t}^{-1}(\alpha) \quad (2.2)$$

which implies the main quantile statement:

$$P\left(y_t < q_{y_t}^{(\alpha)}\right) = \alpha \quad (2.3)$$

The prediction interval I_t^{α} is defined as $(1-\alpha)$-confidence interval into which observation y_t is expected to fall with probability $1-\alpha$. Therefore, it can be described as a range satisfying

$$P(y_t \in I_t^{\alpha}) = P(y_t \in [L_t^{\alpha}, U_t^{\alpha}]) = 1 - \alpha \quad (2.4)$$

where L_t^{α} and U_t^{α} are, respectively, the lower and upper bound of prediction interval I_t^{α} defined by the corresponding distribution quantiles as:

$$L_t^{\alpha} = q_{y_t}^{\alpha_l=(\alpha/2)} = F_{y_t}^{-1}(\alpha/2) \quad (2.5)$$

$$U_t^{\alpha} = q_{y_t}^{\alpha_u=(1-\alpha/2)} = F_{y_t}^{-1}(1-\alpha/2) \qquad (2.6)$$

For instance, with $\alpha = 0.05$, the prediction interval has a 95% confidence level bounded by quantiles $L_t^{0.05} = q_{y_t}^{0.025}$ and $U_t^{0.05} = q_{y_t}^{0.975}$ as $\alpha_l = 0.025$ and $\alpha_u = 0.975$. The above equations are also expected to be correct for the estimations $\hat{f}_{y_t}$ and $\hat{F}_{y_t}$ that are provided by a probabilistic forecaster. The corresponding quantiles for the predictive distributions would hence be $\hat{L}_t^{\alpha}$ and $\hat{U}_t^{\alpha}$ [66][86].

In practice, $\hat{f}_{y_t}$ has to be predicted at time $\acute{t} = t - k$ using all of the information available at time $\acute{t}$ and the probabilistic forecaster would accordingly provide $\hat{I}_t^{\alpha}$ as the prediction interval for the target value in *k* temporal steps (e.g. hours) ahead.

2.2. Forecast Uncertainty Modeling

When a forecasting system is not functioning based on fitting probability models, output forecasts have no guidance about their accuracy. In this situation, the uncertainty dynamics of the predictions have to be analyzed in a secondary procedure [7][16][47]. This is a condition holding for the WRF NWP forecasts which are originally deterministic.

The *ensemble* execution approach is the common approach applied in this situation for weather forecasting. A number of different NWP setups (e.g. 50) with various initial conditions and/or parameters are run as the ensemble members for the same location and horizon in the future. In fact, this approach is a Monte-Carlo method to approximate the stochastic dynamics of the NWP system due to uncertainties in the model and initial conditions [86][72] (Figure 2.2).

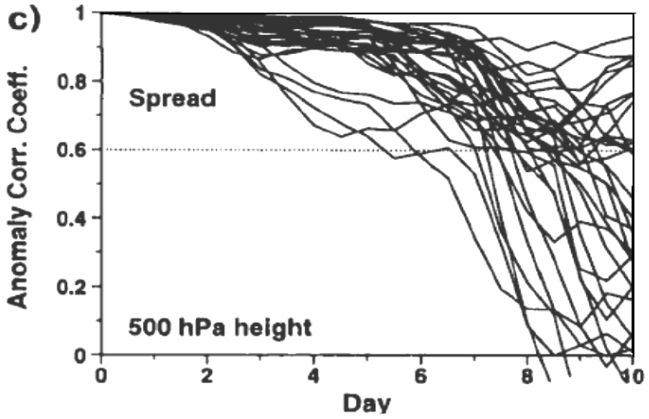


Figure 2.2. Anomaly correlation of 500 hPa height over Europe for individual ensemble members as a function of days ahead. From Molteni *et al.* (1996)

As a result, when the forecast values of the ensemble members have a wide spread the forecast is estimated to be more uncertain compared to situations where there is lower dispersion among the members. To obtain each ensemble member a single run of an NWP model is needed. This incurs a considerable amount of computational cost when running multiple members in an ensemble approach and specifically for very short-term forecasts (e.g. few hours) the members will not have enough time to adjust and spread from their initial perturbations.

It is known that the extent of forecast uncertainty varies with the weather situation [62]. For example, low pressure systems are known to be less predictable than the generally more stable high pressure systems. The historical performance of an NWP system is a valuable source of information about such patterns and can be efficiently used in a post-processing method to model the behavior of the system's forecast uncertainty.

Lange *et al.* [48][49] use a historical performance dataset of wind speed predictions to study the uncertainty in different meteorological situations. The authors cluster wind speed, wind direction, and pressure data into six separate classes of meteorological conditions. The characteristics of these clusters are analyzed to ascertain practical dissimilarities in their forecast uncertainties [47]. The results [48][49] confirm that uncertainty in the forecasted wind speed depends on the forecast weather situation. However, the results of this analysis are not practically considered as a method of obtaining conditional PIs for wind speed forecasts.

A practical application of weather classification to obtain PIs is proposed in Pinson *et al.* [68][69] where two predicted variables, wind speed and wind power, are used to categorize the situations into four manually defined classes. For each class, or situation, the error distribution of its members from the past forecasts would be different. The error distribution of a new forecast case is then expected to follow the distribution of past forecasts with the most similar situations. The distribution of error for a new forecast is thus constructed by bootstrap sampling. The fraction of samples drawn from each class is determined based on the fuzzy membership of the new case in that class. The prediction interval of a new case is then computed based on this reconstructed estimation of error distribution using the empirical statistical quantiles. Therefore, the PIs are practically computed by the categorization of prediction conditions. These resulting intervals are then analyzed and evaluated in a set of experiments. The main shortcoming of this approach is that the classification of the forecast conditions is performed manually by simply discretizing the variables into equal bins (two bins for wind speed and three bins

for wind power). Such simple manual grouping of the prediction outputs does not provide optimal grouping of data points that have high similarity to each other within a group. Hence, the quality of the computed PI is not optimal either. This issue will be even more significant when a larger number of variables/features may have influence on forecast uncertainty. Please note that more detailed background study is provided in each of the individual following chapters as various chapters focus on different areas of uncertainty modeling.

In Chapter 4, we propose more advanced methods both to provide optimal clustering of the historical performance data set and to learn explicit quantile functions. Yet, in the current section, the general process of modeling uncertainty from accuracy records is elaborated.

The systematic characterization of forecast error can simply lead us to the modeling of forecast uncertainty for the target variable. This can be achieved by considering e_t in Equation (2.1) as an instance of the random variable e and associating F_t^e (or its estimation $\hat{F}_t^e$) as its cumulative distribution function, i.e.,

$$\hat{L}_t^{\alpha} = \hat{y}_t + \hat{q}_{e,t}^{(\alpha/2)}, \quad \hat{q}_{e,t}^{(\alpha/2)} = \hat{F}_t^{e^{-1}}(\alpha/2) \quad (2.7)$$

$$\hat{U}_t^{\alpha} = \hat{y}_t + \hat{q}_{e,t}^{(1-\alpha/2)}, \quad \hat{q}_{e,t}^{(1-\alpha/2)} = \hat{F}_t^{e^{-1}}(1-\alpha/2) \quad (2.8)$$

where $\hat{q}_{e,t}^{(\alpha)}$ is the estimated α quantile of "error" based on the estimated forecast error distribution $\hat{f}_t^e$. The distribution of y_t, and hence the desired quantiles, are not explicitly known. Therefore, to find the $\hat{I}_t^{\alpha}$ prediction interval of y_t, the quantiles of e (i.e., the error associated with the forecast) are estimated and added to the predicted value $\hat{y}_t$ to obtain the lower and upper bounds for the original variable [68]. Thus, by finding quantiles over the forecast error distribution, one can find the quantiles over the forecast value that is expected to enclose the target observation.

2.3. Forecast Evaluation

The evaluation of PI forecasts and generally probabilistic forecasts is a more complex process compared to point forecasts [57]. To empirically put our proposed approaches into test, we apply the developed PI models into two real-world data sets. Detailed background and discussions on PI forecast evaluation is provided in Chapter 3. In that chapter we also develop a comprehensive evaluation framework that covers the major measures from the PI evaluation literature and also brings some new insight to the PI

verification process leading to more reliable judgments. This framework is extensively used in the experimental study of different models in the following chapters.

Chapter 3
An Evaluation Framework for Prediction Interval Forecasts

Various measures for the assessment of prediction interval models are discussed in this chapter. Also arguments on the evaluation of the skill of interval forecasts are provided. An uncertainty bound is proposed to incorporate sampling uncertainty in the measurement of forecast skill score.

3.1. Basic Verification Measures

Generally, prediction intervals are better understood and easier to use compared to a full probabilistic distribution function (PDF) of a predicted variable [17][86]. However, the probabilistic nature of interval forecasts complicates their verification compared to deterministic forecasts. The verification of PI forecasts can determine the quality of a forecaster and lead us to proper selection or modification of a forecasting system. Atmospheric science has been the field with most developments in forecast verification processes among others [86]. However, verification analysis of probabilistic and PI forecasts in particular are still under ongoing development [9][12][86]. The Brier score is a classical and still widely used verification score for probabilistic forecasts [8]. However, this score is appropriate for dichotomous variables [39][65]. An extended version of the Brier score for multi-category probabilistic forecasts is the Ranked Probability Score (RPS) [56] which is widely in use. Yet, the inappropriateness of this score is a known fact as it does not penalize vague forecasts i.e. wide intervals [86]. Another version of this score developed for probabilistic forecasts of continues variables named Continuous Ranked Probability Score (CRPS) [33] can only provide a measure for the match between the forecast full PDF and that of the observation [86]. Hence, these

measures are not appropriate for PI evaluation either since they are not sensible to interval width or do not match the double quantile format of PI forecasts.

Other analytical tools such as rank histograms [2] are commonly used for the examination of ensemble forecasts but do not provide a quantitative measure for objective verification of PIs [65]. Recently, information theory approaches have also been studied as a verification tool for probabilistic forecasts. Although the basic idea of using the logarithm of probabilities was proposed in [27], an information theory verification measure named "ignorance score" was first developed in [73]. This score measures the joint entropy of the forecasted probabilistic distribution with the observed distribution. More recently, a new alternative information theoretic measure is proposed in [65]. This measure, called the "information gain", alleviates the tendency of the previous ignorance and CRPS scores to infinity. Although all of these measures can provide proper scores [9][26] for continuous variable probabilistic predictions they are not widely used [12]. One reason for this can be the low intuitiveness and low power of these scores for communicating the forecast skill to the decision makers. It should also be mentioned that these scores can be applied for the verification scenarios where both the predicted and target probabilities are provided in the form of a full probability distribution. Hence, they cannot be directly employed for the evaluation of prediction interval forecasting systems. Instead, efforts have been made to quantify the skill of prediction interval forecasts using more relevant measures. Specific measures for evaluation of "reliability", "sharpness" and "resolution" aspects of PI forecasters have been proposed in [7][67][68]. These measures reflect individual quality aspects of PI forecasters and do not provide a single score for a concluding verification.

The major expectancy from a set of PI forecasts is that their empirical coverage of the observations in a test setting is as close as possible to the required confidence level. This primary property of a PI forecaster *M* is called "reliability" noted as Rel_M^{α}. [68]:

$$\xi_i^I = \begin{cases} 1 & if\ \hat{L}_{\hat{y}_t}^{\alpha} \le y_t \le \hat{U}_{\hat{y}_t}^{\alpha} \\ 0 & otherwise, \end{cases} \quad (3.1)$$

$$Rel_M^{\alpha} = \bar{\xi}_M^{I^{\alpha}} - (1-\alpha) \quad where \quad \bar{\xi}_M^{I^{\alpha}} = \frac{1}{T}\sum_{i=1}^{T} \xi_i^{I^{\alpha}} \quad (3.2)$$

where T is the number of PIs in the test data set used for the evaluation of PI forecasts and ξ_i^I is an indicator of hit if the observation is within the PI boundaries, otherwise it will express a miss by being set to zero. Hence, Rel_M^{α} simply accounts for the difference

between average hit of the forecasts over the whole test cases and the required nominal coverage defined for the PI. For an ideal case we should have $Rel_M^\alpha = 0$ when $\bar{\xi}_M^{I^\alpha} = 1 - \alpha$. Note that we assume, without loss of generality, that all of the forecasts in the tests are provided with a constant confidence level.

A forecasting system that provides PIs with less vagueness expressed by the width of a PI is clearly preferred. This is due to the fact that lower uncertainty in the PI forecasts would lead to a higher value for the exploitation of the predictions. This leads to the second major aspect of PI forecast quality called "sharpness" [59][68]:

$$Width_i^\alpha = \hat{U}_{\hat{y}_i}^\alpha - \hat{L}_{\hat{y}_i}^\alpha \quad (3.3)$$

$$Shp_M^\alpha = \overline{Width}_M^\alpha = \frac{1}{T}\sum_{i=1}^{T} Width_i^\alpha \quad (3.4)$$

where $Width_i^\alpha$ is the width of the i^{th} prediction interval. Note that the sharpness measure has a negative orientation as we prefer forecasts with lower values of average PI width.

Another important quality aspect of a PI computation method is its ability to provide intervals of variable width, depending on the forecast situation. A method with high "resolution" Res_T^α is capable of distinguishing low uncertainty forecasts versus high uncertainty ones and assign wider intervals to forecasts with high uncertainty and narrower intervals to forecasts with low uncertainty. The standard deviation of PI widths is a natural choice to measure the method's resolution [66]:

$$Res_M^\alpha = \left[\frac{1}{T-1}\sum_{j=1}^{T}\left(\hat{U}_j^\alpha - \hat{L}_j^\alpha - Shp_M^\alpha\right)^2\right]^{\frac{1}{2}} \quad (3.5)$$

It should be noted that the resolution measure is not dependent on the observations. Thus, it can be hedged and is not a significant measure individually. However, when the two first major measures of reliability and sharpness are equal for two competing methods, the one with higher resolution may be preferred.

3.2. Forecast Skill Measure

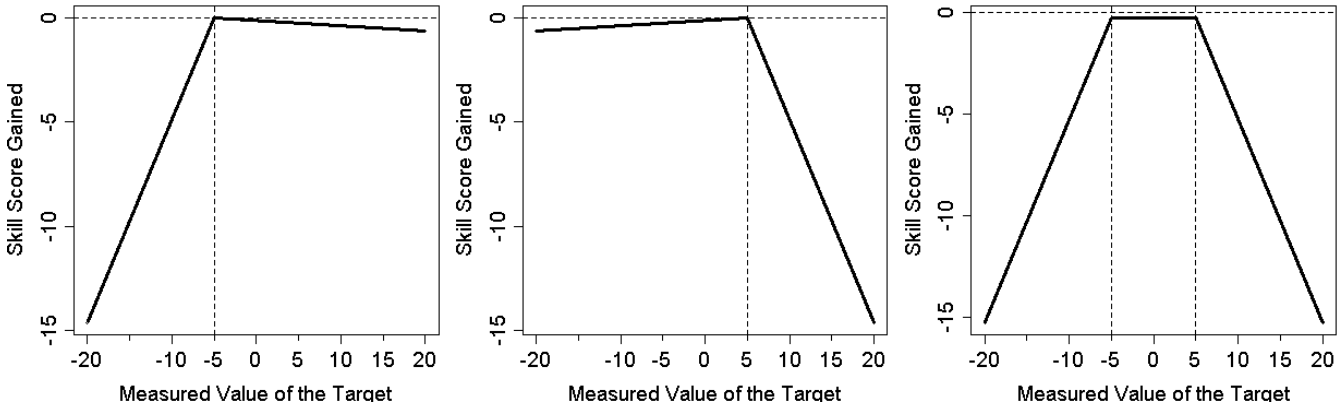
Having a single scalar summary measure of forecast quality is always attractive and useful for objective comparison of various methods, as any such measure would simplify the evaluation of the complete performance profile of a forecaster. The most common prediction interval skill score is the Winkler's score proposed in [87] and is widely used as the concluding objective evaluation measure for PI forecasting methods including in [6][59] and [69]. A comprehensive study done by Gneiting and Raftery [26] proved that this score is "strictly proper" and would hence give the maximum score to a forecast that

is actually the true belief of the forecaster and cannot be "hedged". This would mean that only a PI that follows the true distribution of the target can obtain the maximum score.

Using the notations and assumptions defined here, this skill score can be expressed as:

$$\xi_i^q = \begin{cases} 1 & if\ y_t \le q \\ 0 & otherwise \end{cases} \quad (3.6)$$

$$SScore_M = \sum_{i=1}^{T} \left[\left(\xi_i^{\hat{L}_{\hat{y}_i}^{\alpha}} - (\alpha/2) \right) \left(y_i - \hat{L}_{\hat{y}_i}^{\alpha} \right) + \left(\xi_i^{\hat{U}_{\hat{y}_i}^{\alpha}} - (1 - \alpha/2) \right) \left(y_i - \hat{U}_{\hat{y}_i}^{\alpha} \right) \right] \quad (3.7)$$



(a) (b) (c)

Figure 3.1. Skill score over different observations achieved by a PI of [-5, +5]-conf = 95% for (a) the lower quantile, (b) the upper quantile, (c) the whole PI

Note that the minus of the same objective is minimized as a loss function (error) in the applications of quantile regression for prediction interval analysis, using only one of the terms in the brackets as each quantile is modeled separately [6][59]. To better understand the behavior of this score we algebraically simplify it by considering cases of hits and misses. $\hat{L}_i$ and $\hat{U}_i$ are used instead of $\hat{L}_{\hat{y}_i}^{\alpha}$ and $\hat{U}_{\hat{y}_i}^{\alpha}$ for simplicity. When a "hit" happens for forecast PI of case *i*, we have $\xi_i^{\hat{L}} = 0$ and $\xi_i^{\hat{U}} = 1$. By substituting these values in (13) and multiplying the terms we have:

$$SScore_{i\,(hit)} = -\frac{\alpha}{2} y_i + \frac{\alpha}{2} \hat{L}_i + \frac{\alpha}{2} y_i - \frac{\alpha}{2} \hat{U}_i = -\frac{\alpha}{2} \left(\hat{U}_i - \hat{L}_i \right) = -\frac{\alpha}{2} Width_i^{\alpha} \quad (3.8)$$

In the other case, when an observation is "missed", it is either on the right or the left side of the area outside the PI boundaries. In this case, the values of ($\xi_i^{\hat{L}}$ and $\xi_i^{\hat{U}}$) will be equal to (0,0) or (1,1), respectively. When the missed observation is on the right side of the interval it would have a positive distance of δ_i from the upper boundary $\hat{U}_i$, the score of this particular case can be calculated by using Equation (3.7):

$$SScore_{i\,(right-miss)} = \big(0-(\alpha/2)\big)\big(\widehat{U}_i+\delta_i-\hat{L}_i\big)+(0-(1-\alpha/2))\big(\widehat{U}_i+\delta_i-\widehat{U}_i\big)$$
$$= -\frac{\alpha}{2}\widehat{U}_i-\frac{\alpha}{2}\delta_i++\frac{\alpha}{2}\hat{L}_i-\delta_i+\frac{\alpha}{2}\delta_i=-\frac{\alpha}{2}\big(\widehat{U}_i-\hat{L}_i\big)-\delta_i=-\frac{\alpha}{2}Width_i^{\alpha}-\delta_i \quad (3.9)$$

For a miss happening on the left side of PI, an equal score will also be gained as calculated by Equation (3.9). As $(1-\bar{\xi}_M^{I^{\alpha}})$ will be the overall miss rate, the total score gained by a PI forecasting method *M* over the whole *T* cases in the test set will be:

$$SScore_M = T\left(-\frac{\alpha}{2}\overline{Width}_M^{\alpha}-(1-\bar{\xi}_M^{I^{\alpha}})\bar{\delta}_M^{\alpha}\right)=-T\left(\frac{\alpha}{2}\overline{Width}_M^{\alpha}+\bar{\Delta}_M^{\alpha}\right) \quad (3.10)$$

where $\bar{\delta}_M^{\alpha}$ is the average distance of an observation from the PI boundaries among the missed cases and $\bar{\Delta}_M^{\alpha}$ is the average of this distance among all of the test cases owing to the fact that $\bar{\Delta}_i^{\alpha}$ is equal to zero for hit cases and δ_i for miss cases.

Figure 3.1 depicts the value of this score for different observation values for a sample PI of [-5, +5]-conf=95%. As can be seen the value of the score will be equal for any case where the observation is inside the PI. The score linearly decreases as the observation gets far from the PI boundaries. By multiplying a $2/\alpha$ term to this score, which does not change the actual comparison among different methods, the equal verification measures of Winkler's score and Gneiting's score can be easily retrieved. Hence, it is shown here that all of these scores are essentially equivalent and simply use a weighted sum of the two major aspects of PI quality, namely sharpness and reliability, for verification. However, the reliability aspect of the PI forecaster is here measured by the distance of observations to boundaries ($\bar{\Delta}_M^{\alpha}$).

It is shown in the next subsection that this score must be accompanied by an uncertainty analysis on its own evaluation as with limited number of test samples (which naturally happen in real-world scenarios) it can give misleading results.

3.3. Uncertainty of Skill Score Measurements

The described PI skill score can provide a measure for evaluation of a group of PI forecasters. This way, the best forecaster can be chosen for the application at hand among many potential choices.

Due to the limited availability of test samples, such skill score measurements are subject to sampling variations. Therefore, it is crucial to assess whether the observed skill is due to chance, or if it is a true attribute of the forecasting system. Joliliffe and Stephenson [39] point out: "It has been unusual in weather forecast verification studies for any attempt to be made to assess this sampling uncertainty, although without some

such attempt it is not possible to be sure those apparent differences in skill are real and not just due to random fluctuations". In this work, by decomposition and statistical analysis of the skill score measurements, we consider such variations in the evaluations and hence offering a much more reliable and fair comparison for the user.

A close look at the empirical evaluation study conducted in this work reveals this issue as the mere calculation of the *SScore* measure using a test data set results in misleading evaluations. Since the number of real-world test cases is always limited, the "measurement uncertainty" of *SScore* using the available test data set must also be accounted for. This issue is of a greater importance when there are fewer test cases available to measure *SScore* in each cluster as the number of clusters increases.

To analyze the uncertainty of the skill score, we take a closer look at its components. The terms that are dependent on the method under verification are $\overline{Width}_{\mathrm{M}}^{\alpha}$ and $\bar{\Delta}_M^{\alpha}$. These terms are essentially a weighted sum of their measured values in the *K* different clusters:

$$\widehat{SScore}_M = -T\left(\frac{\alpha}{2}\overline{Width}_M^{\alpha} + \bar{\Delta}_M^{\alpha}\right) = -\sum_{j=1}^{K}|T_j|\left(\frac{\alpha}{2}\overline{Width}_M^{\alpha,j} + \bar{\Delta}_M^{\alpha,j}\right) \quad (3.11)$$

where T_j is the set of test cases that are assigned to cluster *j*. The measured *SScore* is denoted as $\widehat{SScore}_M$, since it is a sample statistic from a single sample set only. $\overline{Width}_M^{\alpha,j}$ is calculated as the average of PI widths among these test cases in cluster *j*. As the PIs in a single cluster are obtained based on the same fitted error distribution, it follows that

$$Width_i^{\alpha,j} = \overline{Width}_M^{\alpha,j} \;\; \forall i \in T_j \quad (3.12)$$

Hence, the width term of the skill score is constant in each cluster, and there is no uncertainty when this statistic is measured using a sample data set in model test evaluations. However, the $\bar{\Delta}_M^{\alpha,j}$ term is the mean statistic of the random variable $\{\Delta_i^{\alpha,j} \mid i \in T_j\}$ which is measured using a set of sample values with T_j members, and thus it is subject to sampling variations.

With limited number of test cases and high nominal coverage rates of PIs, it may happen that in cluster *j=1* only a few test cases (e.g. $|T_1| = 400$) are assigned to a cluster and a fewer number of them (e.g. 30) would lead to non-zero values of $\Delta_i^{\alpha,1}$. The measured value of $\bar{\Delta}_M^{\alpha,1}$ for this cluster may be equal to $\bar{\Delta}_M^{\alpha,2}$ of another cluster *j=2* which has significantly more test cases (e.g. $|T_2| = 6000$). Although these two statistics are

equal, the uncertainty of $\bar{\Delta}_M^{\alpha,2}$ is much smaller than the uncertainty of $\bar{\Delta}_M^{\alpha,1}$ since for cluster two, the measure has been evaluated using a much larger sample set.

To analyze the uncertainty of $\widehat{SScore}_{\mathrm{M}}$, $\bar{\Delta}_M^{\alpha,j}$ must be considered not as a single estimate over the test cases, but as a one-sided confidence interval that provides an upper bound over this measure with a specific confidence level. After using this upper limit for all clusters, a lower limit on the $SScore_M$ with the desired confidence level can be determined:

$$P\left(\bar{\Delta}_M^{\alpha,j} < \bar{\Delta}_M^{\alpha,j^{\beta}}\right) = \beta \quad (3.13)$$

$$P(SScore_M > SScore_M^{\beta}) = \beta \quad (3.14)$$

where β is the desired confidence level over the measure expressed as a percentage. As an example for $\beta = 0.95$, $SScore_M^{0.95}$ is the lower boundary of which the true skill score of method M is expected to be at least equal to with a 95% confidence.

To find the confidence interval over $\bar{\Delta}_M^{\alpha,j}$, its sampling distribution (providing the probability distribution that describes the batch-to-batch variations of this statistic) has to be considered. Bootstrap resampling is a method for building a collection of artificial data batches with the same size as the original sample set with replacement [88].

An example of a sampling distribution of $\bar{\Delta}_M^{\alpha,7}$ and its confidence bound for a sample cluster of test cases for a quantile regression model using spline-basis functions (this method is described in Chapter 5) is shown in Figure 3.2.

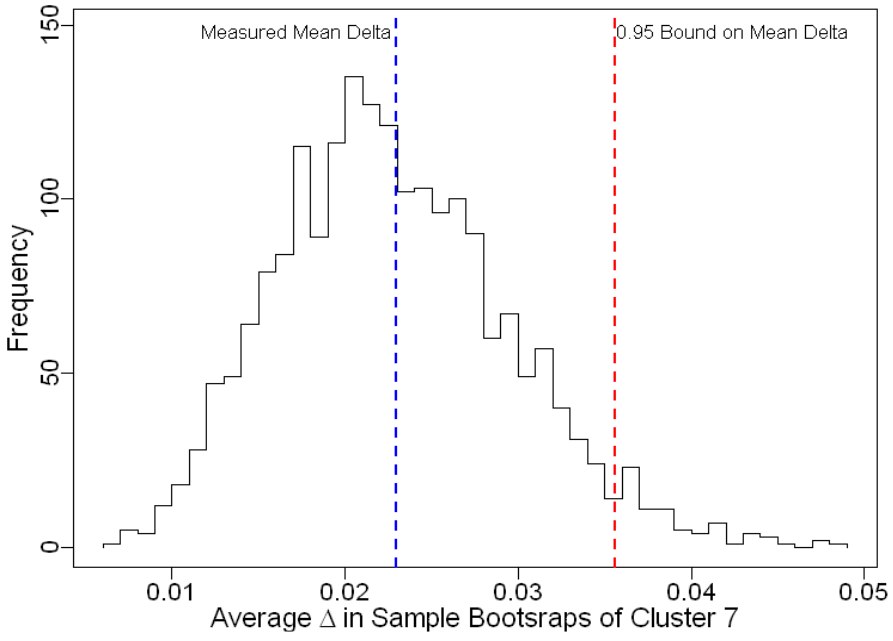


Figure 3.2. Bootstrap distribution of average delta for a sample cluster - #test cases=588, #misses=26

The computed statistic over these batches effectively provides an estimation of the sampling distribution [86]. For the purpose of this study, as many as 2000 bootstrap samples were constructed for each cluster and the $\widehat{\bar{\Delta}_M^{\alpha,J}}$ measure was calculated for each sample set. The distributions defined over these measurements are then used to compute the desired quantile based on the confidence level β. Intuitively, there should be less uncertainty associated with $\widehat{\bar{\Delta}_M^{\alpha,J}}$ when increasing number of test cases in cluster *j* are used in the bootstrapping process. Using the upper limit $\bar{\Delta}_M^{\alpha,j^{\beta}}$ in Equation (3.11) leads to the lower limit of the final skill score in $SScore\,_M^{\beta}$. This measure, which considers the test sample uncertainties, is preferred for fair verification of PI forecasters.

3.4. Other Related Measures and Statistical Tests

From the statistical point of view, the modeled error distribution can be evaluated in comparison to the observed error. Goodness-of-fit tests can assess the hypothesis that the observed data has been drawn from a reference probability distribution. In the PI computations problem, the reference distribution is the fitted error distribution in the training model and the observed forecast errors can be compared to this hypothesized distribution.

Chi-Square Test. The Chi-Square test is a common goodness-of-fit test that compares the frequencies of points in discrete classes of the probability distribution and is hence more appropriate for discrete random variables. When applied to a continuous variable, the data has to be assigned into discrete bins. The test statistic compares the expected and observed frequencies in each of the classes [86]:

$$\chi^2 = \sum_{bins} \frac{(\#Observed - \#Expected)^2}{\#Expected} \quad (3.15)$$

With a good fit, the actual number of observed samples in a bin is very close to the expected number of samples based on the fitted probability distribution. The test statistic would have a sampling distribution of the χ^2 distribution with degrees of freedom equal to (#bins-#fit parameters-1). This is obtained under the null hypothesis that the observed data were drawn from the originally fitted distribution [86].

Kolmogorov-Smirnov (K-S) Test. Unlike the chi-squared test which compares the pdfs of the fit and observation, the K-S Test focuses on the cdfs of these data samples. This test is more proper for continuous variables. The test statistic would be:

$$D = max_x|F_{obs}(x) - F(x)| \quad (3.16)$$

where $F_{obs}(x)$ is the empirical cumulative density function of the observations (refer to Equation (4.8)) and $F(x)$ is the theoretical reference cdf of choice. Hence *D* is essentially the largest difference between the empirical and fitted cdfs over any possible value of *x*. In order to assess the rejection of the null hypothesis the critical values of the *D* statistic are obtained from a table usually constructed by statistical simulations.

It should be noted that the goodness-of-fit tests do not provide quantitative verification scores for forecasts in the form of prediction intervals and can only be applied to full probabilistic distribution cases. Yet, they are a powerful tool to statistically analyze the agreement between the observational and estimated full distributions that are modeled within the PI computation process. Such tests have been adapted and used for the evaluation of conditional coverage in PI forecasts [18]. However, Pinson et al. [69] argue that due to the temporal and persistent nature of weather the independence assumption of PIs would be loose and hence the significance levels by statistical tests can be misleading.

In addition to the various measures discussed above, there have been efforts made to develop probabilistic forecast evaluation scores that are more comprehensible and plausible for a non-expert audience such as [28].

Chapter 4
Clustering Approaches to Weather Forecast Uncertainty Modeling

This chapter proposes the application of clustering methods in the context of NWP forecast uncertainty modeling. Various clustering methods including Fuzzy C-means and density estimation methods work in combination to learn prediction interval models from performance history. Experiments practically apply the proposed models in a realistic set-up and measure the quality and accuracy of the resulting interval forecasts.

4.1. Introduction

With the notion of the dependence of forecast error on the forecast situation [62] in mind, a fine grouping of situations can lead to clusters of forecast cases with a similar error behavior. Simultaneously, the error behavior in a cluster would be distinct from cases in other clusters. Such groupings can be found by clustering all available cases using the relative influential variables as the features. Subsequently, the prediction interval analysis described in the previous section can be applied to each cluster separately. This way, different PIs can be found for the different discovered forecast situations. In other words, rather than considering all cases as equal, the error distribution within each cluster determines the prediction interval of that cluster only. Characteristics of the error distribution for each cluster are found using the past performance of the forecasting system in that cluster (which represents a weather situation) only.

The steps of this PI computation process and their generic input and output are depicted in Figure 4.1. In the first step, the forecast history undergoes the clustering process and the result will be clusters of forecast records. Note that these clusters are determined by the influential features only and the forecast errors are not used in this phase. In the next step, the recorded errors of forecasts in each cluster are modeled by

density fitting algorithms which would provide the density models. Finally, these density models are utilized to calculate the desired prediction intervals that can be used in any application and would also undergo the evaluation and skill analysis process.

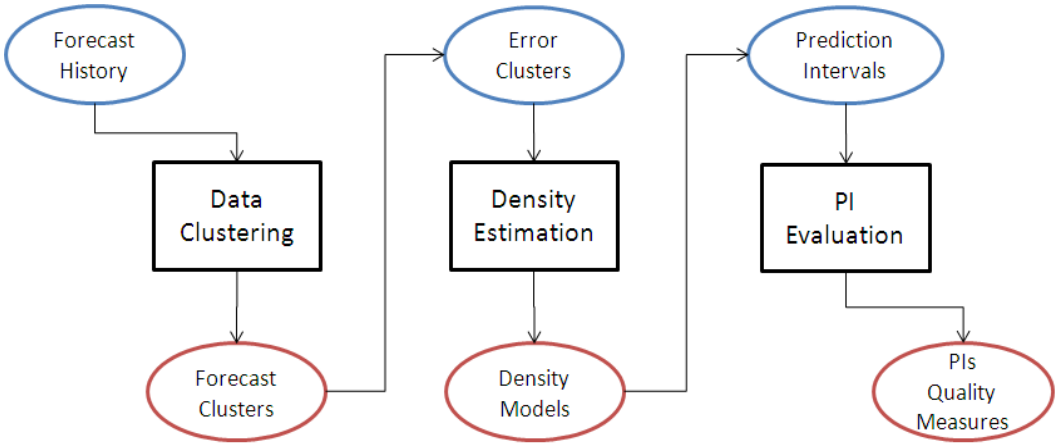


Figure 4.1. The process of uncertainty modeling for PI computation and evaluation

In this study, we apply four different clustering algorithms for the grouping of the NWP past forecasts. Each of these algorithms is one of the most widespread algorithms from a distinct class of clustering algorithms [89] i.e. K-means from median-based algorithms, CLARA from medoid-based algorithms and Agglomerative Hierarchical clustering from hierarchical algorithms. The fourth algorithm is the Fuzzy C-means clustering which is described in section 4.4.

4.2. Fitting Distributions to Forecast Error

Gaussian Fit. The error of a point forecast at time t (e_t) can be regarded as a sample of the error random variable *e*. This random variable *e* would have its own probability distribution which can be characterized by its bias (μ_e) and standard deviation (σ_e). Let $\{e_t\}$ be a series of random samples of the error variable *e*. Then, the values of sample bias and sample standard deviation can be calculated by the following sample statistics:

$$\hat{\mu}_e = \frac{1}{N}\sum_{t=1}^{N} e_t \quad (4.1)$$

$$\hat{\sigma}_e = \left[\frac{1}{N-1}\sum_{t=1}^{N}(e_t - \hat{\mu}_e)^2\right]^{\frac{1}{2}} \quad (4.2)$$

where N is the size of the sample series [17]. A simple yet popular method to find the boundaries of $\hat{I}_t^{\alpha}$ is based on the assumption that the error ($\hat{f}_t^{e}$) follows a Gaussian distribution. Many studies do confirm that the forecast error of many weather attributes

follow a Gaussian distribution [46][47][48]. For a Normal distribution $N(\mu_e, {\sigma_e}^2)$ with known parameters one can calculate the PI quantiles [40] in Equations (2.7) and (2.8):

$$\hat{U}_t^{\alpha} = \hat{y}_t + \mu_e - z_{\alpha/2}.\sigma_e \quad (4.3)$$

$$\hat{L}_t^{\alpha} = \hat{y}_t + \mu_e + z_{1-\alpha/2}.\sigma_e \quad (4.4)$$

where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are the quantiles with $(\alpha/2)$ and $(1-\alpha/2)$ of the standard normal distribution $\mathcal{N}(0,1)$, respectively. In the case of a PI of 95% ($\alpha = 0.05$), $z_{\alpha/2}$ and $z_{1-\alpha/2}$ are equal to 1.96 [29][40]. This method is parametric, as it assumes a normal distribution for the error. Figure 2.1 shows the error distribution of temperature forecasts in various locations in the province of British Columbia (BC), Canada for Summer 2008. The matching normal distribution and the quantiles for the 95% intervals ($\alpha = 0.05$) are also shown.

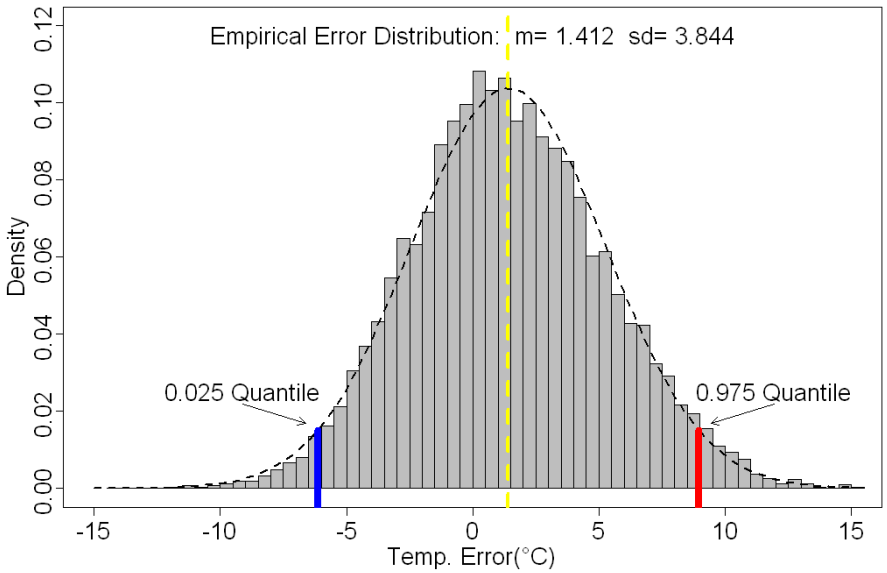


Figure 4.2. Temperature error distribution and corresponding normal distribution based on μ^e and σ^e of the entire available dataset.

Depending on the prediction method in use, different ways to estimate the error distribution parameters in (4.3) and (4.4) have been proposed [26]. Some of them have been shown to be unjustified (e.g., [43]), and many others are not applicable here as our prediction model is NWP and not a time series learning model [17]. An alternative is to estimate the error distribution and its quantiles using a dataset of past performance of the forecast model. Prediction interval calculation methods of this type use the *observed* distribution of errors in the historical records of the system. They are known to provide reasonably good results when theoretical formulas cannot be applied [17]. To obtain PIs

for the NWP forecasts using this method, a dataset of past predictions and associated observations must first be constructed. Subsequently, the prediction interval can be determined by first fitting a Gaussian distribution to the data (i.e. $\{e_t\}$) and then using Equations (4.3) and (4.4) to find the quantiles. By using the simple method of moments for the fitting step one can apply Equations (4.1) and (4.2) to calculate $\hat{\mu}_e$ and $\hat{\sigma}_e$ using sample statistics of the empirical dataset. Hence, as these parameters are estimates, the boundaries of the prediction interval are determined using the following equations [40][88]:

$$\hat{L}_t^{\alpha} = \hat{y}_t + \hat{\mu}_e - t(\alpha/2\,, N-1).\hat{\sigma}_e.(\tfrac{1}{N}+1)^{1/2} \quad (4.5)$$

$$\hat{U}_t^{\alpha} = \hat{y}_t + \hat{\mu}_e + t((1-\alpha/2)\,, N-1).\hat{\sigma}_e.(\tfrac{1}{N}+1)^{1/2} \quad (4.6)$$

where $t(\alpha, n)$ is the quantile of the Student's *t*-distribution for confidence level α and n degrees of freedom. The quantiles of *t*-distribution and the multiplier term are used since the moments of the real distribution are unknown. Rather, they are estimated based on samples from the historical performance dataset [40].

Weibull Fit. Investigations into actual forecasts accuracies show that in many cases the forecast error distribution does not fully follow a symmetrical normal distribution shape. This is often viewed in target attributes that follow non-Gaussian distributions such as wind speed [66]. To achieve a better fit and consequently better PIs for such cases with skewness, Weibull distribution can be a potentially proper choice. Weibull distribution has two parameters *k* and λ as follows:

$$f(x;\lambda,k) = \left(\frac{k}{\lambda}\right)\left(\frac{x}{\lambda}\right)^{k-1}\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right] \quad (4.7)$$

where $k > 0$ is the shape parameter and $\lambda > 0$ is the scale parameter of the distribution and the value of function is zero for $x < 0$. To find the fitting distribution parameters for a set of x values, the method of Maximum Likelihood Estimation (MLE) is used [86]. Using MLE the distribution parameters are tuned into values that the expectation of drawing the sample data from the fitted distribution would be maximized.

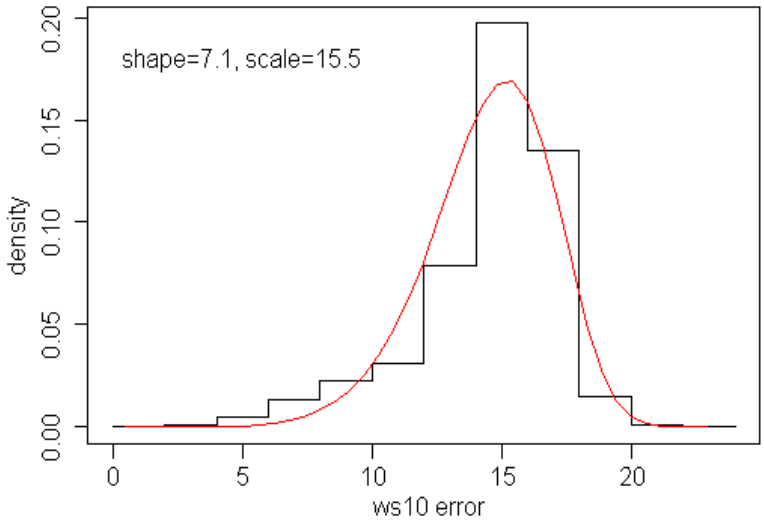


Figure 4.3. Wind speed (m/s) error distribution and a its Weibull distribution fit (curve) for a sample subset of NWP forecasts

The cdf of this fitted Weibull distribution ($\hat{F}_t^e$) can then be used to compute the error quantiles of the PI when used in Equations (2.7) and (2.8). It must be noted here that the error values in the fitting process have to be shifted to right so that the minimum value for the random variable would be zero. This can be done by adding the magnitude of the worst possible error for the target to all entries of the error set. Figure 4.3 depicts an example of fitting Weibull over a set of wind speed errors.

Empirical Distribution. Another alternative for the analysis of error distribution is to not assume any predefined type of distribution over the samples. The error distribution and the respective PIs in such cases can be derived from the actual distribution of the sample data at hand. This means that the *empirical* cumulative distribution function of sample errors is used as a direct estimate of the true population distribution [66][68]. This empirical cdf is defined as:

$$\hat{F}_t^e(q) = \frac{1}{N}\#\{e_t \in E | e_t < q\} \quad (4.8)$$

where E is the set of errors in the available sample. By this approach the term $\hat{F}_t^e$ in Equations (2.7) and (2.8) will be computed differently ultimately leading to different PIs. Because there has not been any assumption made on the forecast error distribution, such approach is called a *non-parametric* method of distribution estimation and probabilistic forecasting.

Kernel Density Smoothing. There are some potential drawbacks in the application of the empirical cdf method for estimation of forecast error distribution. First, the sampling characteristics of the data can have a dramatic impact on the cdf function. Second and more significantly, the domain of the pdf function will be limited to the minimum and maximum values existing in the sample which is not ideal for the PI analysis that is enormously sensitive to the tails. An alternative to the empirical pdf function is the kernel density smoothing method which can both provide a smoother function and a better estimate of the tails. Instead of considering a 0 or 1 binary value in the empirical pdf construction, kernel density smoothing is achieved by stacking kernel blocks that are centered at the data values. A smoothing kernel function is a non-negative function that has a unit area and hence is a proper probability density function on its own. In Figure 4.4 four different types of smoothing kernels are shown. The support of these function are [-1,1] except for the Gaussian kernel which has a support of $[-\infty, +\infty]$.

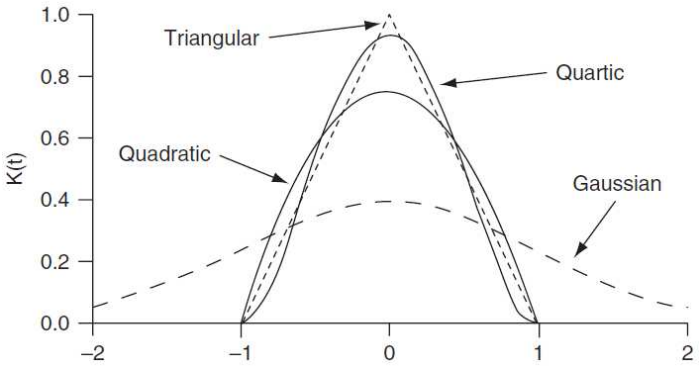


Figure 4.4. Four common smoothing kernels [86]

Each sample would provide a stacking element equal to the smoothing kernel centered at the sample data value and the final pdf function would be constructed by:

$$\hat{f}_t^e(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x-x_i}{h}\right) \quad (4.9)$$

where h is the smoothing parameter which balances the smoothing intensity. A good choice for this parameter is critical and when using the Gaussian kernel a reasonable choice would be [75]:

$$h = \frac{min\left\{0.9s, \frac{2}{3}IQR\right\}}{N^{1/5}} \quad (4.10)$$

where s is the sample standard deviation and IQR is the Inter-Quantile Range of the sample data.

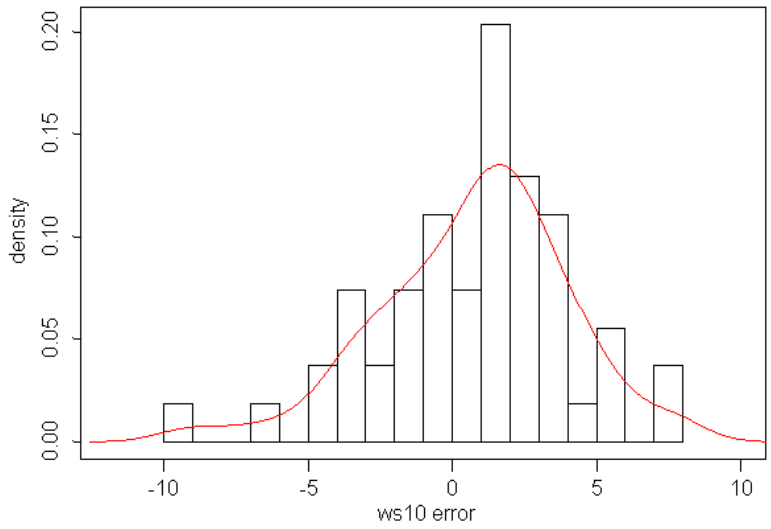


Figure 4.5. Empirical pdf and the kernel smoothing density of a sample wind speed error set

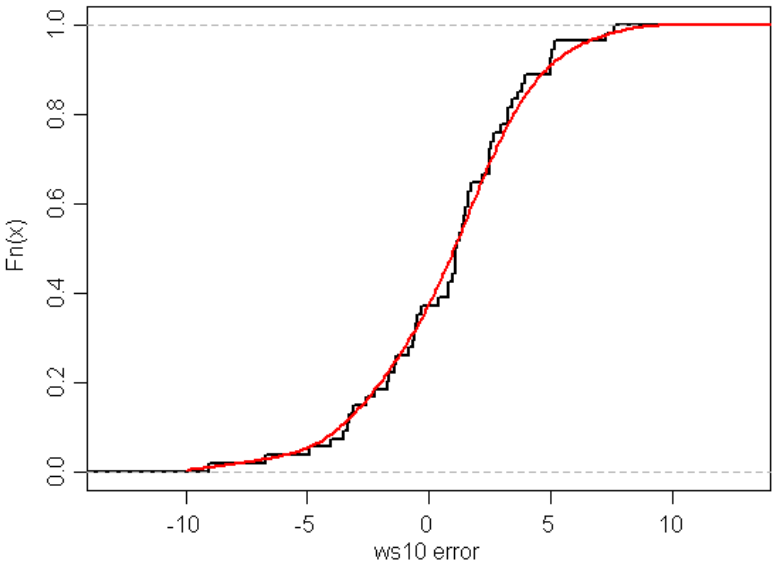


Figure 4.6. Empirical cdf and the kernel smoothing cdf of the sample wind speed error set in Figure 4.5

In Figure 4.5 the empirical distribution of a sample subset of wind speed errors are shown. The curve shows the kernel smoothing density of this sample using Gaussian kernels. Also the cumulative distribution function of the empirical distribution and the kernel smoothing distribution for this sample are shown in Figure 4.6. As can be seen the kernel version desirably has a smoother shape and declines gradually on the edges.

It should be noted that almost always the assumption of “perfect observation” cannot be made. If the observational error is not comparable with the forecast error it can be ignored. Otherwise, a possible solution would be to add a noise (with the variance of

observation error) to the forecasts [2]. In all of the above error fitting procedures, the weather situation and other features (such as time and location) do not play any role. Thus, these PI computation methods cannot yield different intervals in the various weather situations. When applied to the NWP model outputs, each of these methods would estimate a prediction interval for the given attribute (e.g., temperature) with the same width no matter where and when the forecast is made and what the weather situation is.

To change the above parametric and non-parametric error fitting and PI computation methods into a dynamic method and to make them conditional with respect to the weather situation, we propose the use of clustering so that the previous analysis can be applied to different well-distinguishable forecast situations. Later on, other learning methods are studied to provide dynamic PIs from error records.

4.3. Prediction Interval Computation Using Crisp Clustering

In crisp clustering algorithms every data point is strictly assigned to a single cluster and hence the partition matrix of the clustering has binary elements. In this section we elaborate on the crisp clustering algorithms and their application in prediction interval modeling.

K-means. This algorithm is a simple yet powerful clustering algorithm that has been used in many applications [84] including clustering of atmospheric situations and patterns [35]. To find k clusters in a dataset $D = \{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_N\}$ where $\mathrm{x}_j = \{x_j^1, x_j^2, \dots, x_j^d\}$, N is the total number of available forecast cases for training and d is the total number of influential variables. The K-means clustering algorithm iteratively updates the center points of the K clusters $C = \{c_1, c_2, \dots, c_K\}$ and reassigns every data point to the nearest cluster center. This heuristic iterative process locally minimizes the total distance of points to their respective cluster's center [41]:

$$J = \mathrm{argmin}_{\mathrm{C}} \sum_{i=1}^{K} \sum_{\mathrm{x}_j \in D_i} \left\| \mathrm{x}_j - c_i \right\|^2 \quad (4.11)$$

where D_i is the set of points in D that are assigned to cluster i as c_i is their nearest cluster center in C. Here, $x_j^{1..d}$ are the d influential features for forecast case j. We would also have the forecast error of case j associated with the predictand y as e_j^y but not used in the clustering.

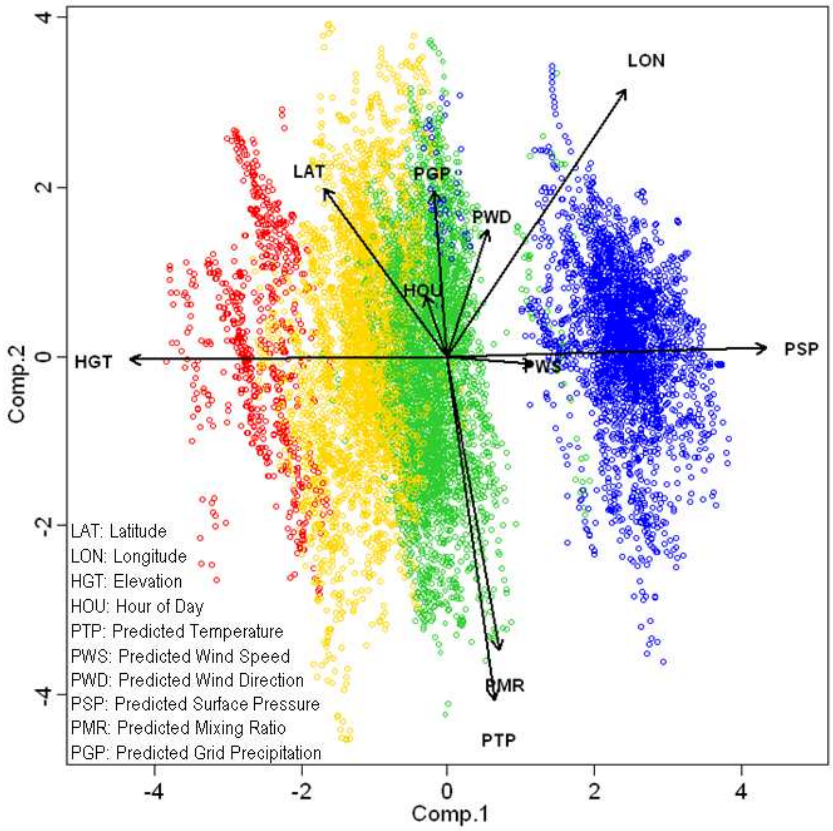


Figure 4.7. 2D PCA visualization of the four identified clusters (cumulative proportion of variance = 0.40).

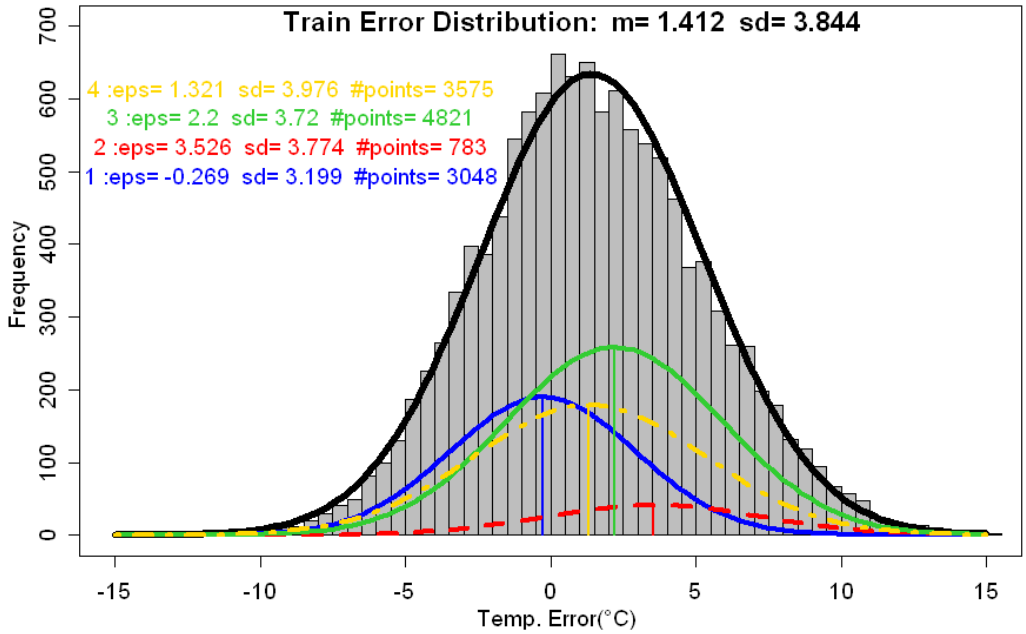


Figure 4.8. Error distributions and their moments for the BC dataset (solid black) and four identified clusters of forecasts.

To find clusters of NWP forecasts, each prediction can be considered a point in *D*. For each point x_i, up to 25 influential variables (d = 25) are taken into account such as:

forecast temperature, wind speed and wind direction (at different geo-potential heights), surface pressure, mixing ratio, grid precipitation, hour of the day of the forecast, and the latitude, longitude, and elevation of the forecast location.

The main thesis of the proposed approach is that clustering of past forecasts can efficiently capture the characteristics of the forecast conditions and categorize them into distinct classes. The forecast error behavior (represented by its distribution) is anticipated to follow the same pattern within a category, but the pattern will differ among categories.

Consequently, after a set of cluster centers $\{c_1, c_2, \dots, c_K\}$ is determined from past forecasts by applying a clustering algorithm, each cluster will have its own set of forecast cases D_i and also its own set of errors for target *y* in E_i^y such that:

$$E_i^y = \left\{ e_j^y \,|\, x_j \in D_i, j = 1..n^i \right\}, i = 1..K \quad (4.12)$$

where n^i is the number of sample points in cluster *i*. The error distribution of a desired variable (e.g., temperature or t2) in each cluster *i* could be determined by considering the past forecasts errors of members in that cluster only. This set is defined as the set E_i^{t2}. Now that the error samples are also grouped based on their forecast situation and the influential variables, any of the distribution fitting approaches described in the previous subsection can be applied on the sets $E_i^{t2}, i = 1..K$.

For instance, based on the Gaussian fitting method each cluster i of forecast errors E_i^{t2} will have its own estimated probability distribution $\hat{f}_{i,t}^{e}$ and sample statistics $\hat{\mu}_e^i$ and $\hat{\sigma}_e^i$. A sample clustering for *K=4* on the BC database is projected into the first two significant components from Principal Component Analysis (PCA) in Figure 4.7. The arrows show the correlations of various attributes used in the clustering process with these components. The Gaussian probability density functions which are fitted to the set of errors from these clusters are plotted along with the original single set of *all* past forecast errors in Figure 4.8.

After the training phase is finished and a new forecast x_{new} is made, the cluster to which it belongs can be identified by the nearest cluster center:

$$\mathrm{x}_{new} \in D_i \text{ , where } i = argmin_j \left\| \mathrm{x}_{new} - c_j \right\|^2 \quad (4.13)$$

and boundaries of the corresponding prediction interval can be estimated:

$$\hat{L}_{new}^{\alpha} = \hat{y}_{new} + \hat{\mu}_e{}^i - t\left(\alpha/2\,, n^i - 1\right) \cdot \hat{\sigma}_e{}^i \cdot \left(\tfrac{1}{n^i} + 1\right)^{1/2} \quad (4.14)$$

$$\hat{U}_{new}^{\alpha} = \hat{y}_{new} + \hat{\mu}_e{}^i + t\left((1-\alpha/2)\,, n^i - 1\right) \cdot \hat{\sigma}_e{}^i \cdot \left(\tfrac{1}{n^i} + 1\right)^{1/2} \quad (4.15)$$

where $\hat{y}_{new}$ is the attribute of interest in the forecast $\mathrm{x}_{\mathrm{new}}$. PIs determined this way would be generally variant between different forecasts as they depend on the cluster of forecast situations to which the current forecast case belongs.

Using the model performance history makes the proposed approach of PI computation efficient and applicable to forecasts obtained from NWP model outputs. To examine the effect of clustering algorithm on the quality of calculated PIs, two other crisp clustering algorithms are considered in this study: clustering for large applications (CLARA) [41], and Agglomerative Hierarchical clustering [36].

CLARA. Random sampling approach is used in the CLARA (Clustering LARge Applications) [41] to handle the large number of points in recent applications such as data mining. The key point is that appropriate sample sizes can effectively maintain the important geometrical properties of the entire data set. To improve the efficiency of the brute force search process in the PAM clustering algorithm [89], CLARA applies PAM to find the representative medoids only in a randomly drawn sample from the data set. For better approximation, CLARA repeats this process with multiple samples from the original data set.

Agglomerative Hierarchical Clustering. In hierarchical clustering the data samples are grouped using either a top-down or bottom-up approach. In the agglomerative case, data rows are regarded as single clusters initially and at each step the most similar pairs of clusters merge into a new single cluster in the higher level and the process is repeated using the new set of clusters [41][89]. Hence, at the end of this bottom-up process we will have a tree structure of clusters (dendrogram) which has the whole data set as a single cluster in the root. Unlike the other two algorithms this algorithm does not require the number of clusters as an input. Instead the resulting dendrogram can be cut into any desired number of clusters based on the order of similarities between the clusters. Further details of these clustering algorithms are available in [89].

4.4. Prediction Interval Computation Using Fuzzy Clustering

The clustering algorithms described in the previous section assign each sample point into a single cluster only. Therefore, the membership of a forecast case in any cluster ($\mathrm{x}_j \in D_i$) is a binary value. In a possibly more natural approach, the forecast cases can be associated with the clusters by different levels of membership. This varying degree of

membership is fundamentally supported by fuzzy set theory [63]. Such partial membership of samples can potentially improve the modeling of forecast situations when analyzing the patterns of forecast history records. In this way, a forecast case can be simultaneously considered as members of various forecast situations. Many weather conditions such as transitional phases of weather can be better explained by this approach.

Fuzzy C-means is one of the most widely used clustering algorithms that can discover cluster patterns based on the fuzzy membership assumption. In this algorithm the membership values of data points (rows) in the various clusters (columns) are represented by a matrix that can have fractional value in the entries rather than only binary values in the case of crisp clustering algorithms. The objective function of the clustering process is changed accordingly [4][64]:

$$J = argmin_C \sum_{i=1}^{N} \sum_{j=1}^{k} u_{ij}^{m} \left\| \mathrm{x}_i - c_j \right\|^2 \quad (4.16)$$

where u_{ij} represents the degree of membership of the point x_i in cluster j and $\sum_{l=1}^{k} u_{il} = 1$. $m > 1$ is the fuzzification factor that controls the balance between membership values of close to 0 or 1 and the values in between.

The objective function can be minimized using Gradient Descent in an iterative process where the membership matrix and the cluster centers will be updated by:

$$u_{ij} = 1 / \sum_{l=1}^{k} \left(\frac{\left\| \mathrm{x}_i - c_j \right\|}{\left\| \mathrm{x}_i - c_l \right\|} \right)^{2/m-1} \quad (4.17)$$

This iterative optimization process will repeat until none of the u matrix entries changes are bigger than ε (convergence).

Finally, these fuzzy patterns of historical forecasts can be used for the modeling of forecast error. Unlike the binary clustering approach that each error sample contributed to a single cluster in the next distribution modeling step, the output of the fuzzy clustering process will determine the contribution of each error sample to all of the *k* clusters. After applying a binary clustering algorithm, the set of past error samples for target *y*, E_i^y could be achieved by Equation (4.12). However, since the samples in the training phase of the clustering process have fuzzy levels of membership for each single cluster, the E_i^y set cannot be determined as before. In addition, any new forecast case x_{new} is now going to be associated to all of the clusters but with different degrees of membership and Equation (4.13) will not be appropriate anymore as the error sample set should be independently

defined for any new forecast case according to its own membership levels. Hence, we should devise a new method to determine the sample error set $E^{y}_{\mathrm{x}_{new}}$ that can describe the forecast error characteristics of each new forecast independently. When $E^{y}_{\mathrm{x}_{new}}$ is made available, any of the distribution fitting approaches described in Subsection 4.2 can be applied to this set in order to provide the estimated forecast error distribution $\hat{f}^{e}$ which can then be used in Equation (2.7) and (2.8) to get the prediction interval quantiles of choice.

The process of determining $E^{y}_{\mathrm{x}_{new}}$ is essentially a probability distribution combination problem. Here we apply the bootstrapping approach which is a resampling method that tries to get a better estimate of a population parameter by measuring the estimate over multiple representative samples [86]. The $E^{y}_{\mathrm{x}_{new}}$ set of errors would have N members (i.e. the number of past sample errors for every new forecast is equal to all available past forecast samples.). Out of these N samples $u_{\mathrm{x}_{new}j}.N$ would be drawn from E^{y}_{j} with replacement. Hence, when $\mathrm{x}_{\mathrm{new}}$ has a higher level of membership in cluster j, more samples from E^{y}_{j} would contribute to $E^{y}_{\mathrm{x}_{\mathrm{new}}}$.

It should also be noted that E^{y}_{j} is a fuzzy set on its own where the vector $u_{ij}, i = 1..N$ determines its members. This implies that the process of sampling $u_{\mathrm{x}_{new}\mathrm{j}}.N$ points from E^{y}_{j} is not uniform and is performed by the weighted probability vector of u_{ij} for cluster j.

Once the described sampling process provides the set $E^{y}_{\mathrm{x}_{\mathrm{new}}}$, the distribution fitting methods can then be applied just as in the binary clustering case. The fitted probability density function $\hat{f}^{e}_{x_{new}}$ would then be used to obtain the cumulative distribution function and the desired quantile values as explained in Subsection 2.1. The PI quantiles $\hat{L}^{b,\alpha}_{new}$ and $\hat{U}^{b,\alpha}_{new}$ would be achieved by a single run of the bootstrapping process where $b = 1..B$. After repeating the process for B times the final PI would be achieved by:

$$\hat{L}^{\alpha}_{new} = \frac{1}{B}\sum_{b=1}^{B}\hat{L}^{b,\alpha}_{new} \quad (4.18)$$

$$\hat{U}^{\alpha}_{new} = \frac{1}{B}\sum_{b=1}^{B}\hat{U}^{b,\alpha}_{new} \quad (4.19)$$

where B is the number of bootstrap samples and $\hat{I}^{\alpha}_{\mathrm{new}}$ can be finally available for usage and evaluation purposes.

As an alternative to the bootstrapping process to obtain a single fitted error distribution for any new forecast, a two phase process can be used. In the first phase, a distribution can be fit for each cluster involving all training samples, using a weighting scheme based on each sample's degree of membership in that cluster. In other words, the training samples that are more associated with a cluster contribute more to the formation of that cluster's pdf:

$$E_j^y = \left\{ e_i^y,\ m_j\left(e_i^y\right) \mid m_j\left(e_i^y\right) = m_j(x_i) = u_{ij},\ i = 1..N,\ j = 1..k \right\} \quad (4.20)$$

Hence, when applying the kernel density smoothing method to fit a probability distribution over the error set of cluster j, u_{ij} determines the vector of weights for the samples in the fitting process. In the second phase (forecasting), any new forecast case x_{new} is now associated with all clusters, but with different degrees of membership. Therefore, the PI boundaries computed for each cluster's fitted distribution are consolidated using membership level of the new forecast in each of these fuzzy clusters

$$\hat{L}_{new}^{\alpha} = \sum_{j=1}^{k} u_{new,j} \hat{L}_{new}^{j,\alpha}, \quad \hat{U}_{new}^{\alpha} = \sum_{j=1}^{k} u_{new,j} \hat{U}_{new}^{j,\alpha} \quad (4.21)$$

where $u_{new,j} = m_j(x_{new})$ is the membership level of x_{new} in cluster j. This provides a normalized weighted mean of the individual quantiles calculated for each cluster. Hence, for example, when the new forecast belongs to cluster 1 with a much higher degree compared to cluster 2, its prediction interval is determined with much stronger contribution from quantiles of cluster 1 rather than cluster 2. This method is in essence a distribution combination process.

1) **Clustering Phase:** Use the d influential weather attributes from the NWP forecasts to discover k cluster centers using Fuzzy C-means (using equation (7)) over all of the forecast records.
2) **Distribution and Quantile Estimation Phase:** For every cluster $l = 1..k$ consider E_l^y as a fuzzy set of error samples defined by the membership level of their associated forecast in this l^{th} cluster (i.e. u_{il}, i=1..N).
    a. Fit a (parametric or non-parametric) distribution ($\hat{F}_l^e$) to cluster l by using u_{il} as the weight vector over the samples.
    b. Use the estimated $\hat{F}_l^e$ to obtain $\hat{L}_l^\alpha$ and $\hat{U}_l^\alpha$ for the prediction interval with $(1-\alpha)$ level of confidence.
3) **Test Phase:** For new forecast x_{new} determine its level of membership in each cluster $u_{new,j}$ $(j = 1..k)$.
4) The lower quantile of the forecast ($\hat{L}_{new}^\alpha$) is obtained by combining the estimated quantiles based on the membership weights i.e. $\sum_{j=1}^{k} u_{new,j} . \hat{L}_j^\alpha$ (likewise for the upper quantile)

Figure 4.9. Steps of fuzzy prediction interval modeling and forecast

The general steps of the process of training and using a prediction interval forecasting system using fuzzy clustering is provided in Figure 4.9.

4.5. Experimental Results

4.5.1. Data Sets and Method Set-ups

We experimentally evaluate the applicability and performance of the described uncertainty analysis models to obtain PI forecasts from the WRF Numerical Weather Forecasting model [80]. Two different hindcast data sets of hourly predictions are accompanied by the respective observations from weather stations from the National Center for Atmospheric Research (NCAR) data repository. By joining the relevant observation for each forecast and deriving the associated forecast error, the two data sets can be considered as two repositories of the NWP model's historical performance. The WRF v3 simulations were run in three nested grids with resolutions of 10.8 km, 3.6 km and 1.2 km. The outermost domain covered an area of about 15595 km^2 with a 38×38 grid. The nearest grid point to the observation station was assigned as the associated forecast grid.

The first data set is for the forecasts and observations of the summer of 2008 in 60 different weather stations in the province of British Columbia, Canada. This data set (referred to as BC) contains about 13,000 records of forecast history. For this data set, 10 major weather, location and time attributes were used as the influential variables:

predicted wind speed, wind direction, temperature, surface pressure, mixing ratio, grid precipitation, altitude, latitude, longitude, and hour-of-day.

The second data set covers a much longer period of time i.e. three years of 2007, 2008 and 2009 for two stations in BC. This data set (referred as AG) contains about 51,000 records of historical performance. There are a total of 33 features available in this data set as listed in Table 4.1. For both of these data sets the observations are used to derive the forecast error for two meters temperature and ten meters wind speed forecasts and the described PI computation methods are applied to achieve PIs over the forecasted temperature and forecasted wind speed. Please also note that since the features of wind direction and hour of day are periodic, we substitute them with their sine and cosine transformations with the appropriate periods (i.e. 360 and 24, respectively) in the best setups of fuzzy clustering and quantile regression (next chapter).

Table 4.1. Available influential variables for the AG historical data set

t2 (temperature at 2m),	ws10 (wind speed at 10m)	wd10 (wind direction at 10m)
psfc (surface pressure)	pg1/3/6/12 (pressure gradient – current value compared to 1/3/6/12 hours before)	td2 (dew temperature at 2m)
psl (sea-level pressure)	rh2 (relative humidity)	temperature at 950/925/850/700/500 pressure levels
horizontal wind speed at 950/925/850/700/500 pressure levels	vertical wind speed at 950/925/850/700/500 pressure levels	wind direction at 950/925/850/700/500 pressure levels
Hour of day	station	-

Table 4.2. Feature sets used for uncertainty modeling of the BC data set

	Features		
Feature Set	10 Basic Feats	pg1, pg3 pg6	PCA
BF1	●		
BF1PG	●	●	
PG		●	
BF1PGPC4	●	●	●
BF1PGPC6	●	●	●

Table 4.3. Feature sets used for uncertainty modeling of the AG data set

	Features			
Feature Set	10 Basic Feats	Pressure levels Feats.	pg1, pg3 pg6, pg12	PCA
BF1	●			
BF2	●	●		
BF1PG	●		●	
BF2PG	●	●	●	
PG			●	
BF1PC4	●			●
BF2PC4		●		●
BF2PC8		●		●
BF2PC12		●		●
BF1PGPC4	●		●	●
BF1PGPC8	●		●	●
BF2PGPC4	●	●	●	●
BF2PGPC8	●	●	●	●
BF2PGPC12	●	●	●	●

For the BC data set five different subsets of the available features are defined to investigate the role of influential variables and to select the optimal set for PI forecasts. These feature sets are describes in Table 4.2. Note that the features starting with "pg" are new derived temporal features from the forecasts as they represent the gradient of surface pressure between the current forecast and the forecasts made in one, three and six hours ago for the same location. It is expected that these features would provide valuable information about the temporal stability of the forecasted weather for the uncertainty model. Also, Table 4.3 lists the 14 different feature sets used for the AG data set. The feature set which excludes the pressure gradient and pressure level attributes is called the basic set here. As the number of features would be extensively high for some setups such as BF2PG and this can have negative impact on the quality of the clustering algorithm the Principal Component Analysis (PCA) technique is applied on some of the feature sets to use the *C*-most important components in the analysis rather than all of the dimensions. The number after PC in these feature set names represents how many of the first significant components were used.

For the evaluation of the various PI forecasting methods each data set is split into a train set and a test set in every experiment. Only the train set is applied for the modeling phase of the process when the clusters of weather forecast situations are discovered and

their error distributions are modeled. After the training phase, the test data which has never been seen is fed into the model to obtain the PIs. These forecasted PIs are then subject of the verification measures as described in Chapter 3.

The available data sets are split randomly into 5 different folds to perform a 5-fold cross validation process to better evaluate the methods. At each step 4 folds of the data are used for training and the 5^{th} fold (which was not applied in the training process) is used in the test phase. In another evaluation setup the AG data set is split into 3 folds based on the temporal sequence of the records. That is at each fold run, two years of data are used for training and the third year of data is used for test. For example the 2007 and 2008 data are used to train the model and then the 2009 data is used to verify the trained PI forecasting model. Monthly split of the data records in the BC data set would also yield to 3-fold cross validation.

Here we focus on the 95%-confidence level PIs (i.e. $\alpha = 0.05$) for temperature and wind speed. Due to the availability of alternative choices for the various steps in the PI training phase many different models will be defined by the combination of these options:

- Feature Sets: as listed in Table 4.2 and Table 4.3.
- Clustering algorithm: K-means (5 random starts), CLARA (300+4*K* random samples) and HClust (Euclidian distance and Ward's agglomeration method), Fuzzy C-means (FCM) (*m*=1.2 for BC and 1.1 for AG)
- Number of clusters (*K*): from 2 to 200 with increasing intervals
- Fitting Method: Gaussian, Weibull, Empirical and Kernel density smoothing

In the test phase the trained model is used to compute the PIs for the forecasts made in the test data set and the resulting PIs are provided as inputs to an evaluation procedure that would determine all of the explained quality measures and scores for the verification of PI forecasts. Each measure is evaluated in every fold test run and its average over all of the folds is considered as the overall estimate of that measure for a method.

To compare the various proposed methods with a baseline method, some simple approaches are considered. The first possible baseline method is the *climatological* approach that considers all of the past error samples together (i.e. *K*=1) and computes the PI based on these samples. Note that any of the fitting methods can be used for its error distribution modeling. Other baseline methods considered in this study would follow manual categorization of past forecast records based on an attribute. Here, we consider

methods that simply use forecast hour-of-day, month and temperature as the categorizing attribute (i.e. K=24, K=12 and K=10).

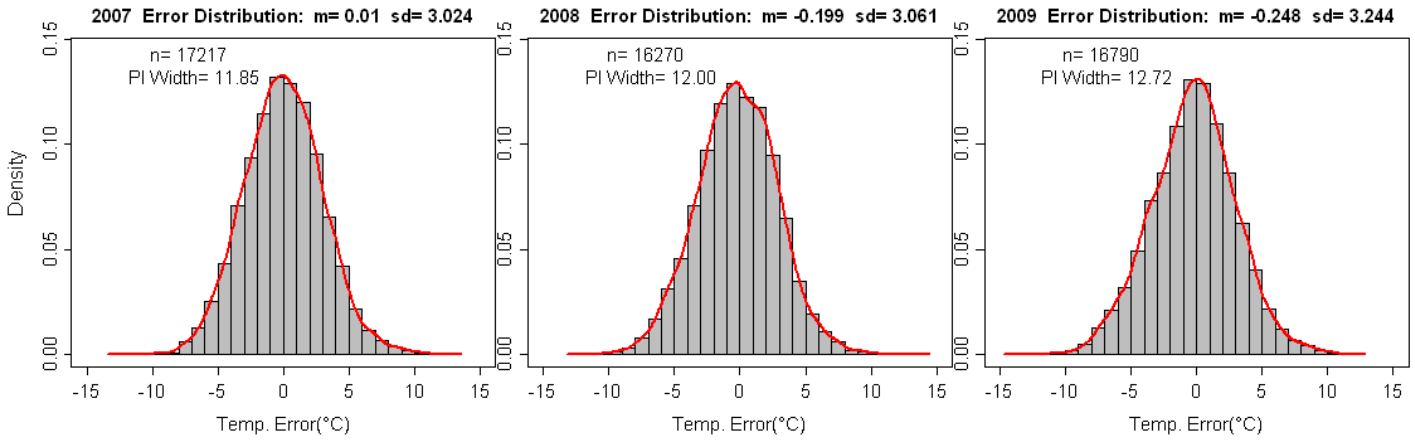


Figure 4.10. Temperature error distributions for various years (in AG data set) with different bias and variances and the width of a Gaussian fit PI

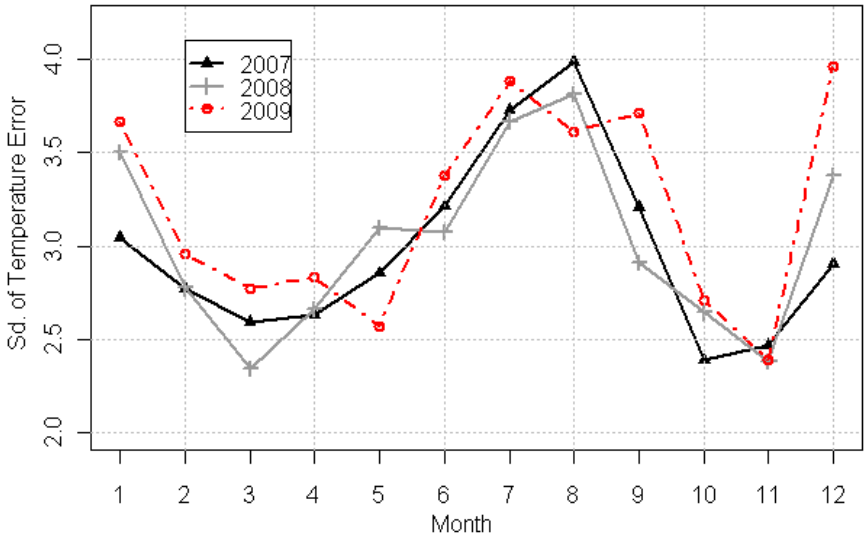


Figure 4.11. Standard deviation of temperature error in each month for different years in the AG data set

To have an initial look at the forecast error attributes in the data Figure 4.8 shows the temperature error distribution in the BC data set. Also in Figure 4.10 and Figure 4.11 the sd. of temperature error (as a key aspect in PI analysis) is plotted for different months and years in the AG data set. As can be seen there clearly is some regular pattern of forecast uncertainty in the different months that can be exploited for obtaining conditional PIs. Similar patterns can be observed for the wind speed forecast errors as well (Figure 4.12).

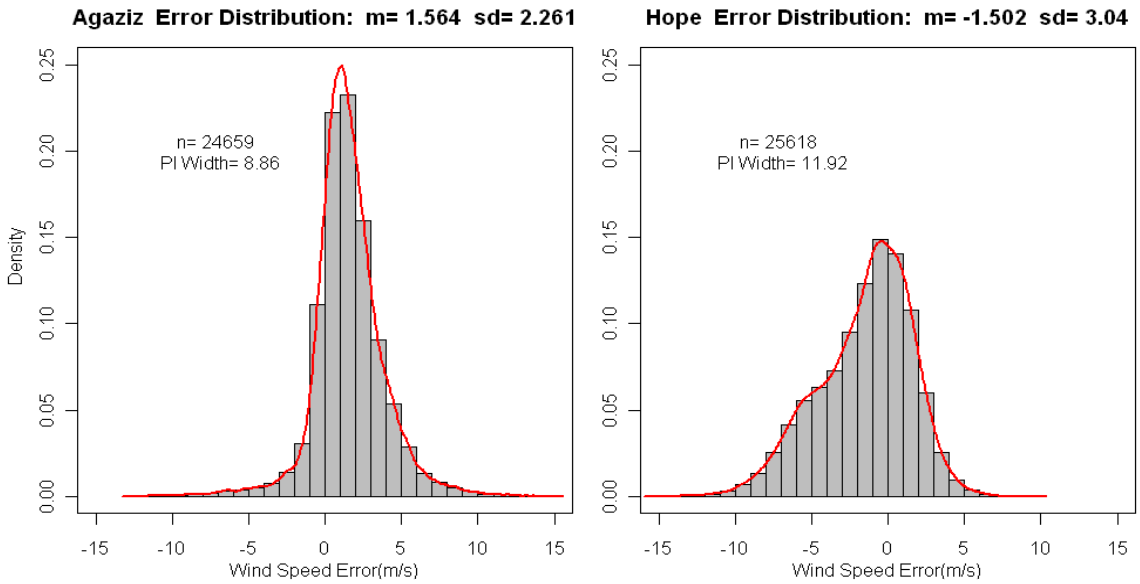


Figure 4.12. Wind speed error distributions for the two different stations (in AG data set) with a notable difference in bias and variances and the width of a Gaussian fit PI

The above analysis confirms the hypothesis that the forecast error behavior would follow different patterns and attributes in various forecast conditions which can be exploited by the PI computation methods to achieve PIs that are dynamic based on these situations.

4.5.2. Crisp Clustering PI Forecasting Methods

The methods that achieved the five best *SScore* for temperature PIs as defined in Equation (3.11) among the entire possible PI forecasting methods (as described in the previous subsection) are listed in Table 4.4. Please note that for simplicity, *SScore* values are divided by $-T$ so the measure is independent from the number of test samples between different experiments and it is also negatively oriented (as in error measures). The best performing methods have used the maximum number of clusters for the clustering process which is counter intuitive as with very large number of clusters there will be very few samples available in each cluster to effectively learn the uncertainty model of the weather situation represented by that cluster. The same issue is evident in the best methods in the AG data set with the yearly cross validation experiments as reported in Table 4.5 and also in the wind speed error PI methods (not reported here).

Table 4.4. Top five methods and the detailed measures for temp. in BC data set based on SScore in 5-fold cross validation

Algorithm	K	Fit	Features	Sharpness	Coverage	Resolution	SScore	SScore Rank
K-means	200	Kernel	BF1	11.45	95.79	3.12	0.3233	1
Clara	200	Kernel	BF1	11.61	95.45	3.08	0.3302	2
HClust	200	Kernel	BF1	11.30	95.04	3.24	0.3359	3
K-means	150	Kernel	BF1	11.77	95.73	2.92	0.3366	4
K-means	200	Empirical	BF1	9.84	90.00	2.82	0.3372	5

Table 4.5. Top five methods and the detailed measures for temp. in AG data set based on SScore in 3-fold (yearly) cross validation

Algorithm	K	Fit	Features	Sharpness	Coverage	Resolution	SScore	SScore Rank
K-means	200	Kernel	BF1	9.99	94.02	1.94	0.3125	1
K-means	150	Kernel	BF1	10.16	94.24	1.83	0.3145	2
K-means	200	Kernel	BF2PG	9.97	93.92	2.26	0.3161	3
K-means	200	Normal	BF1	9.64	93.12	1.89	0.3164	4
K-means	100	Kernel	BF1	10.58	94.81	1.78	0.3171	5

To have a closer look at the role of *K* in the SScore evaluations the trend of this score with increasing number of clusters is provided in Figure 4.13 and Figure 4.14. These figures show the *SScore* trend for the best temperature PI setup using each of the clustering algorithms for the BC and AG data set. The ever improving trend is in contradiction with the statistical nature of the training process.

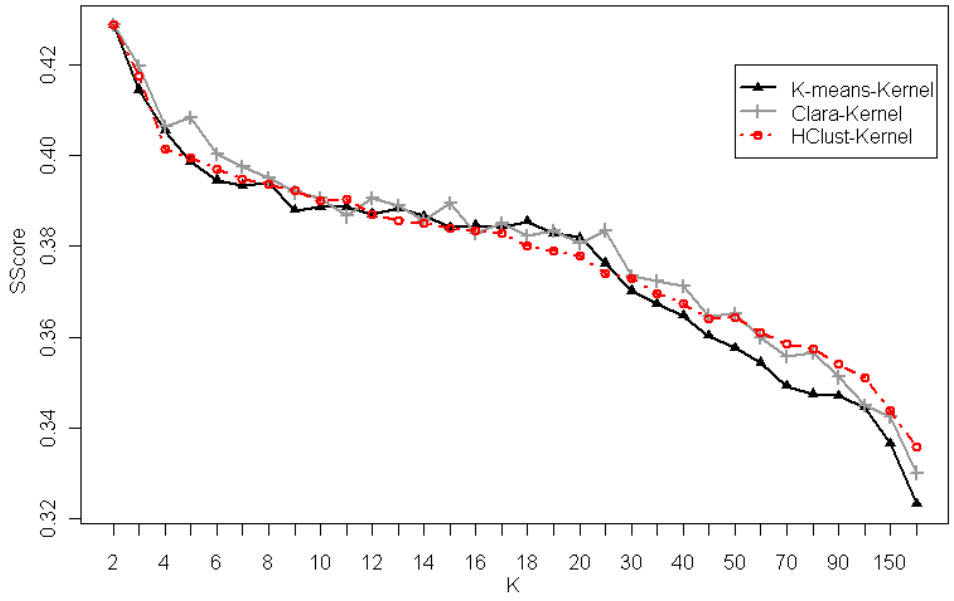


Figure 4.13. *SScore* trend of best temperature PI methods over increasing number of clusters in the BC data set

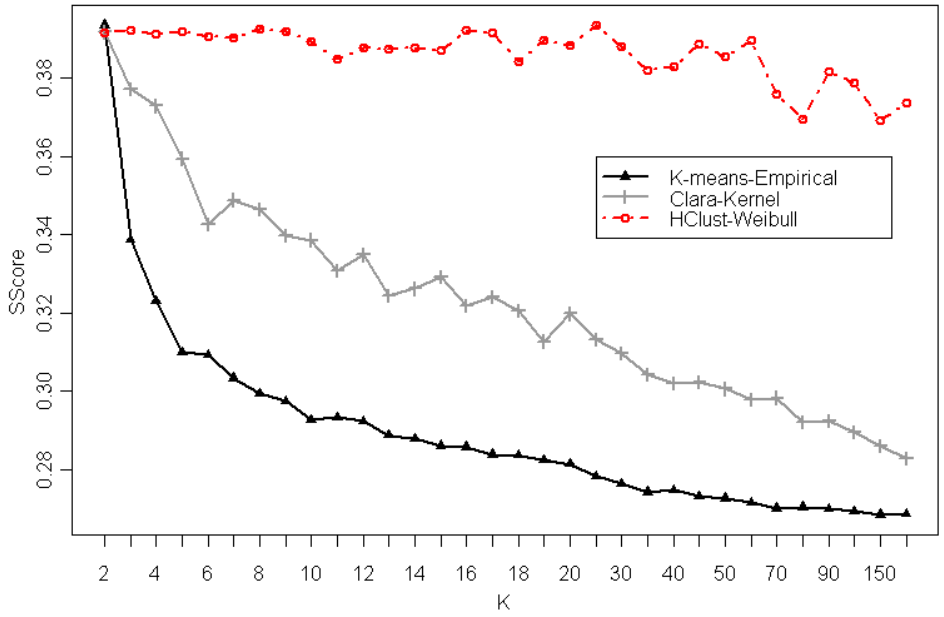


Figure 4.14. *SScore* trend of best temperature PI methods over increasing number of clusters in the AG data set

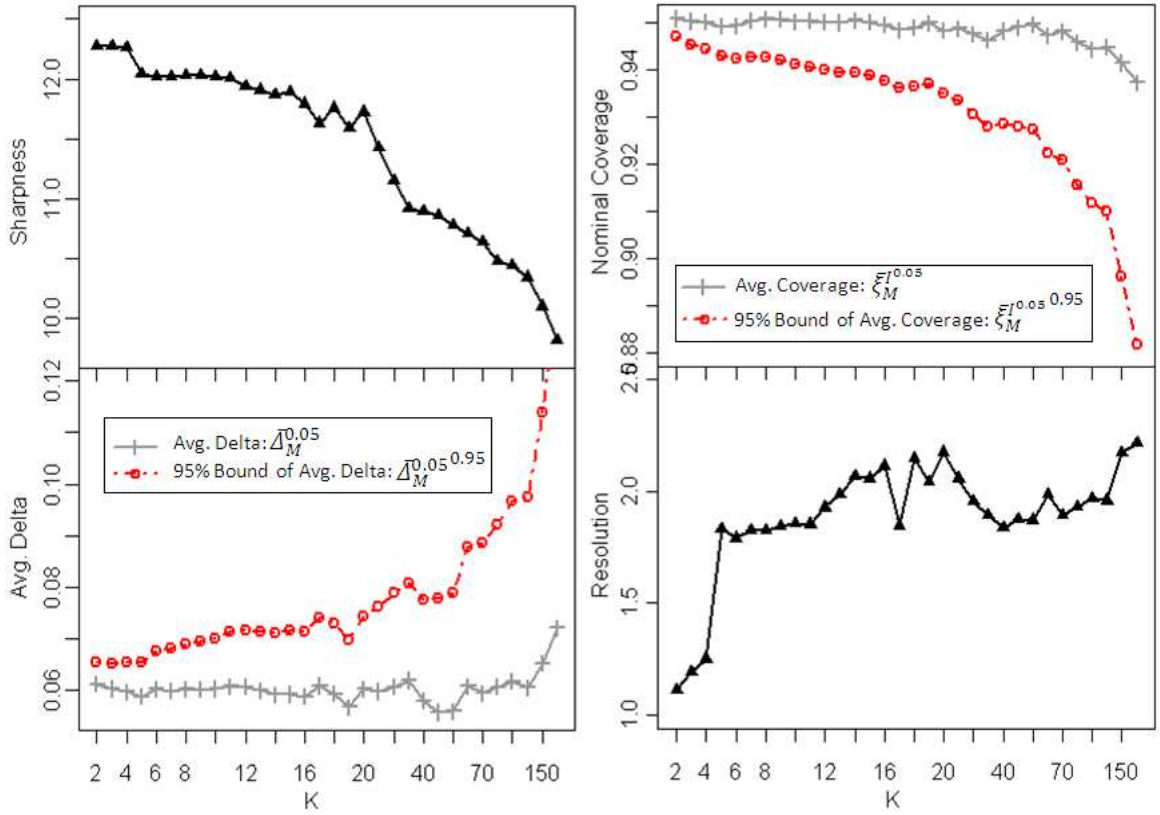


Figure 4.15. The trend of detailed forecasted PI quality measures with increasing number of clusters

The reason for this optimistic measurement of *SScore* in methods with higher number of clusters is that the measured *SScore* is a sample statistic over the available test

samples. Yet, as argued in subsection 3.3, the uncertainty bounds for this measure will be further away from this estimation with small number of samples. By using the 95%-confidence level bound for the *SScore* verification score as defined in (3.14) we can make sure that the judgments are not mislead by the low number of available test samples. As Figure 4.15 shows sharpness and the average delta (as the comprising elements of the skill score) have improving trends when the number of clusters increases. However, the 95% bound on the average delta measure does not follow such trend. The observed nominal coverage, its 95% bound and the resolution measures are also depicted in this figure. Figure 4.16 and Figure 4.17 show the trend of $SScore^{0.95}$ (that uses the delta bound rather than its sample value) for the best setups of the clustering PI methods. As can be seen, this score accounts for the uncertainty of the *SScore* measurement using the available test set and shows a decrease in forecast skill (increase of *SScore*) with large number of clusters as expected.

Consequently, $SScore^{0.95}$ is used to rank the various methods and the results are reported in Table 4.6 and Table 4.7. Also, the best ranks achieved by the various baseline methods are listed in these tables. In the BC data set, the K-means clustering algorithm with 6 clusters and kernel density estimation provides PI forecasts with the average width of 13.42 while the Base-Temp baseline method would provide PIs for the same forecasts with an average width of 14.23. The paired-t test over the skill scores of these two methods confirms that their estimated means are statistically significantly different (p-value=0.014<0.05). The difference between the skill score of the best clustering and the best baseline method is also statistically significant (p-value=0.0001). In addition, the K-means algorithm PIs would have a standard deviation of 1.20 degrees in the forecasts, while this value is equal to zero for the climatological baseline approach that provides constant width PIs.

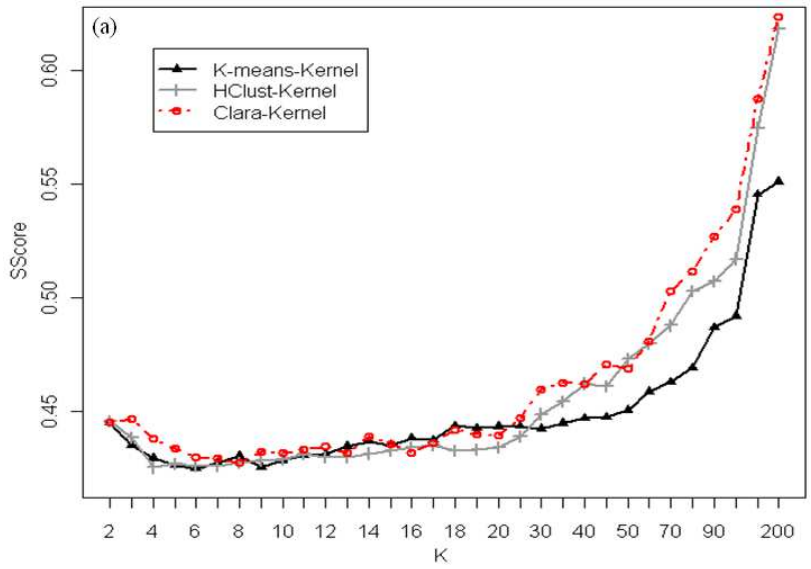
Also in the AG data set, the best K-means setup with the kernel fitting method over the BF2 feature set can achieve an *SScore* which is less than 0.3485 with a 95% confidence. This value would be equal to 0.3774 for the climatological baseline and 0.3704 for the best baseline which is the month-based grouping method (p-value<0.005). Such improvement in forecast skill is achieved as the PIs of the K-means setup have less width (less vagueness) and higher coverage of observations (reliability).

Table 4.6. Top five methods and the detailed measures for temp. in BC data set based on $SScore^{0.95}$ in 5-fold cross validation

Algorithm	K	Fit	Features	Sharpness	Coverage	$Coverage^{0.95}$	Resolution	RMSE	SScore	SScore Rank	$SScore^{0.95}$	$SScore^{0.95}$ Rank
K-means	6	Kernel	BF1	13.42	95.37	93.05	1.20	3.41	0.3946	1034	0.4245	1
HClust	4	Kernel	BF1	13.79	95.66	93.86	0.73	3.40	0.4014	1194	0.4251	2
K-means	9	Kernel	BF1	13.26	95.58	92.68	1.29	3.37	0.3879	818	0.4253	3
K-means	10	Kernel	BF1PGPC4	13.12	95.73	92.60	1.33	3.32	0.3837	610	0.4254	4
K-means	7	Kernel	BF1PGPC4	13.48	95.52	92.96	0.95	3.39	0.3928	998	0.4256	5
Base-Temp.	10	Kernel	Temp.	14.23	95.59	92.59	1.10	3.65	0.4071	1314	0.4423	451
Base-Clim.	1	Kernel	-	14.99	95.19	94.32	0.00	3.80	0.4394	1895	0.4514	782
Base-Ws.	10	Kernel	Ws.	14.49	95.61	92.89	0.92	3.67	0.4235	1487	0.4621	1001
Base-Hour	24	Kernel	Hour	13.72	95.77	90.21	1.09	3.47	0.3972	1111	0.4679	1098

Table 4.7. Top five methods and the detailed measures for temp. in AG data set based on $Sscore^{0.95}$ in 3-fold (yearly) cross validation

Algorithm	K	Fit	Features	Sharpness	Coverage	$Coverage^{0.95}$	Resolution	RMSE	SScore	SScore Rank	$SScore^{0.95}$	$SScore^{0.95}$ Rank
K-means	50	Kernel	BF2	10.78	94.96	92.74	1.87	2.80	0.3254	56	0.3485	1
K-means	45	Kernel	BF2	10.86	94.89	92.78	1.87	2.83	0.3273	78	0.3492	2
K-means	40	Kernel	BF2	10.89	94.82	92.85	1.84	2.83	0.3303	114	0.3499	3
K-means	50	Kernel	BF2PG	10.94	94.87	92.60	2.20	2.87	0.3281	89	0.3506	4
K-means	70	Kernel	BF2PG	10.64	94.75	91.98	1.95	2.78	0.3226	33	0.3509	5
Base-Month	12	Kernel	Month	12.21	95.12	94.10	1.91	3.12	0.3601	2541	0.3704	1671
Base-Temp.	10	Normal	Temp.	11.70	94.44	93.57	0.98	3.04	0.3620	2809	0.3725	2193
Base-Ws.	10	Kernel	Ws.	12.12	94.91	94.17	1.20	3.12	0.3664	3037	0.3754	2681
Base-Clim.	1	Normal	-	12.17	94.78	94.49	0.00	3.11	0.3740	3985	0.3774	2845



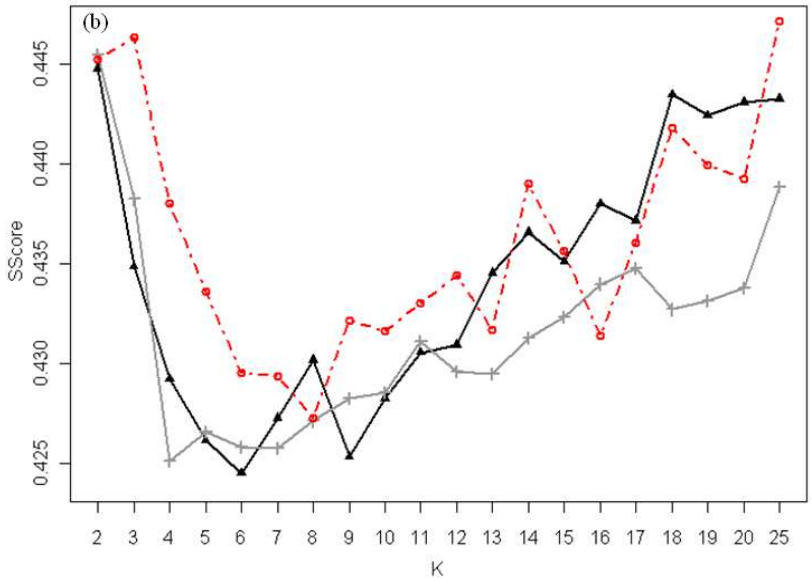


Figure 4.16. $Sscore^{0.95}$ trend of best temperature PI methods over increasing number of clusters in the BC data set for (a) K=2..200 clusters (b) K=2..25 clusters

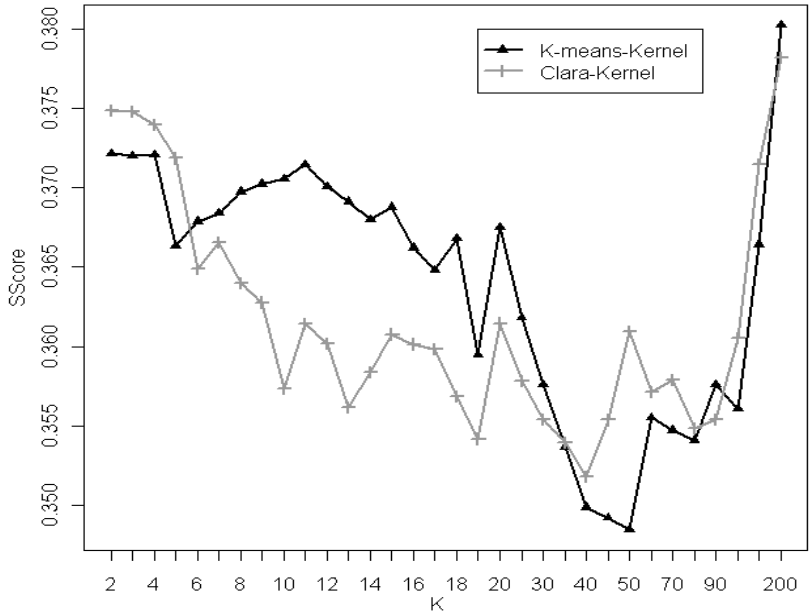


Figure 4.17. $Sscore^{0.95}$ trend of best temperature PI methods over increasing number of clusters in the AG data set

Achieving better scores by bigger number of clusters in the AG data set compared to the BC data set can be a result of the availability of more data samples and features both in the train and test phases as this would increase the complexity of the learning space in the training phase and decrease the uncertainty of the skill score evaluations in the test phase. In the Coverage$^{0.95}$ column of the measure tables, the 95%-confidence level lower bound for the measured nominal coverage is provided. This estimate is the weighted average of the binomial test lower bound of nominal coverage in individual clusters

based on the number of test cases in each cluster. There is a notable difference observed between the sample measure of nominal coverage and its 95%-confidence level lower bound due to the availability of rather few test samples with bigger number of clusters.

Moreover, the Root Mean Squared Error (RMSE) of the forecasts is listed in these tables. This important point forecast performance measure is calculated for the forecasted PIs based on considering the median of the PI as the new revised point forecast. The notable improvement achieved by the proposed methods compared to the baseline methods is due to the dynamic calibration of forecast bias in the forecast groups discovered by the clustering algorithms. Here the forecast bias is estimated from the accuracy records in a dynamic fashion depending on the forecast situation characteristics.

To have a general comparison between the various comprising elements of a PI forecasting method we aggregate and summarize the performance measures of the three clustering algorithms over all of the combinations they can have with number of clusters, feature sets and fitting methods. The same aggregation is performed for the feature set and fitting method elements. Figure 4.18 shows the box plot of skill score of the clustering algorithms for the BC and AG data sets. This figure shows the better performance of the K-means clustering-based PI forecasts in both data sets. Due to the low scalability of the HClust algorithm for the AG data set, a randomly selected subset with half of the size of the original data set was used in the training phase of this algorithm. This seems to be the reason for the reduced skill of HClust PIs in the AG data set.

Figure 4.19 shows the statistics of $SScore^{0.95}$ for the five feature sets in the BC data set. It shows that the BF1 and BF1PGPC4 feature sets achieve the best temperature PI quality. Also comparison of the fitting methods reveals that the Kernel density smoothing method achieves the best scores. For temperature error distribution, the Normal distribution outperforms the Weibull fit but still both methods are not as good as the Empirical method. Very similar results were obtained from the 5-fold cross validation evaluations in the AG data set. The role of AG data feature sets for temperature PIs is investigated in Figure 4.19 and Figure 4.20 and they generally suggest that (lower whiskers reaching to smaller $SScore^{0.95}$ – better skill) the pressure level features (included in BF2) are relevant and helpful for the temperature error modeling and PI computation.

As for the PIs of wind speed forecasts, the clustering algorithms had practically the same comparison as in temperature PIs. Yet, the Weibull fitting method leads to better results here (Figure 4.21). The feature set BF0 is introduced here for wind speed analysis

specifically. After initial results revealed that the simple baseline methods (using only one or no features) performed better that the defined feature sets, a new feature set BF0 was defined that comprises of the predicted wind speed and surface pressure only. As shown in Figure 4.21 and Figure 4.22, BF0 would lead to better wind speed PIs both for BC data set and AG data set. Also the pressure level features in BF2 lead to wind speed PI forecasts with better skill compared to the basic features in BF1.

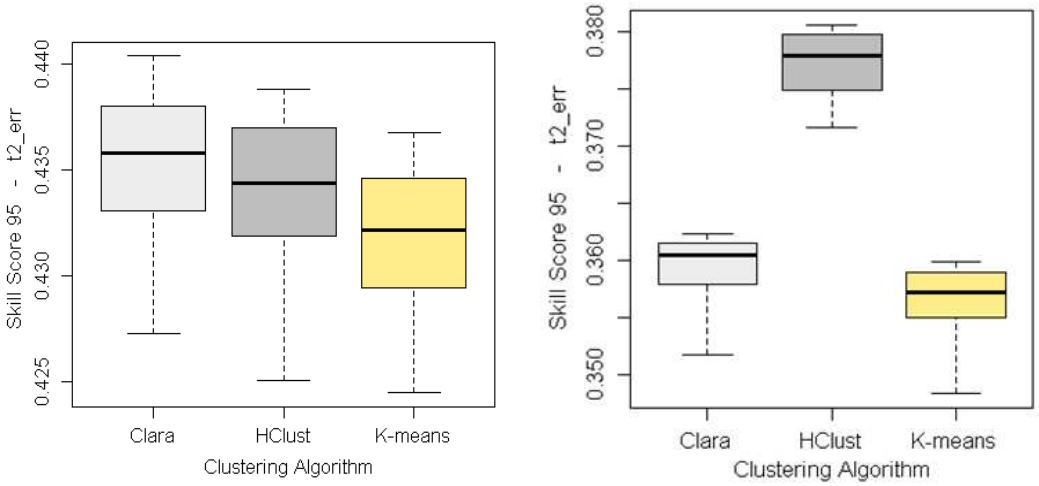


Figure 4.18. $SScore^{0.95}$ of the three clustering algorithm for temperature PIs in (left) BC data set (5-fold cross validation) and (right) AG data set (Yearly cross validation)

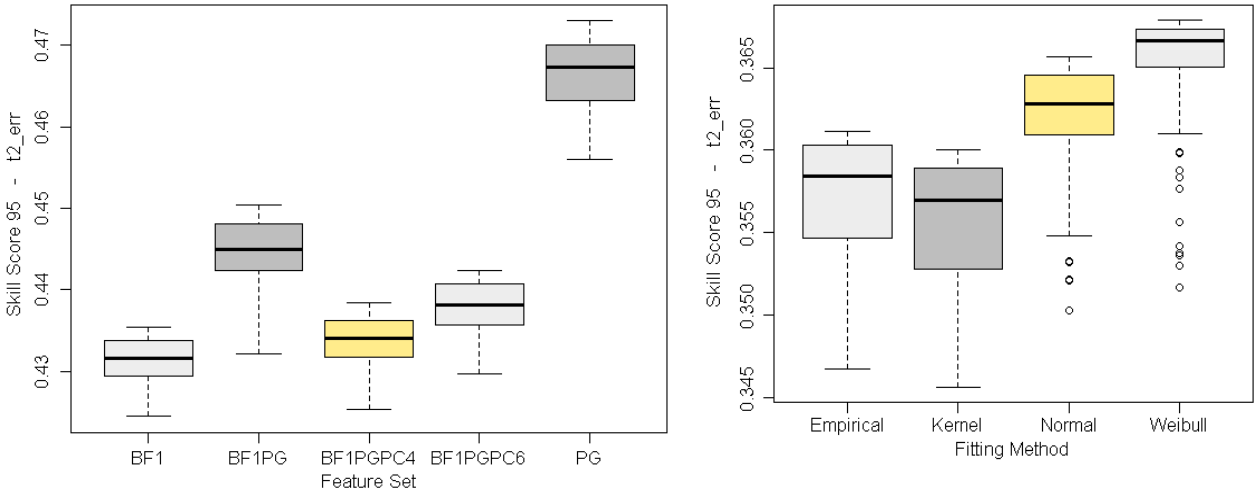


Figure 4.19. $SScore^{0.95}$ of (left) the six different feature sets for temperature PIs in BC data set (5-fold cross validation) and (right) the four different fitting methods in AG data set (Yearly cross validation)

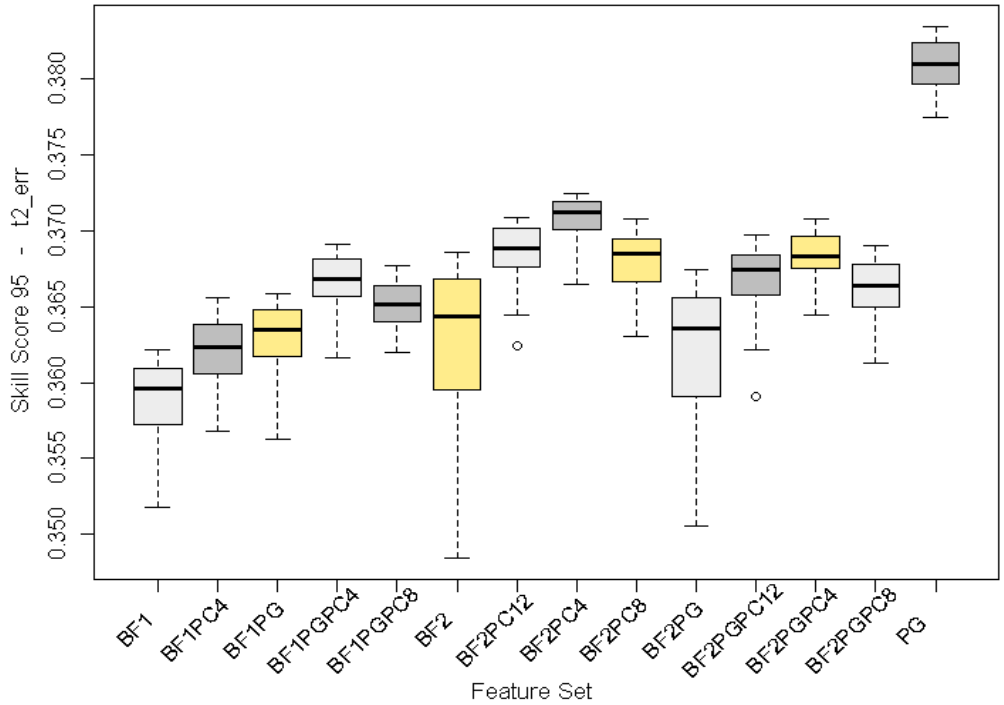


Figure 4.20. SScore$^{0.95}$ of 14 different feature sets for temperature in AG data set (Yearly cross validation)

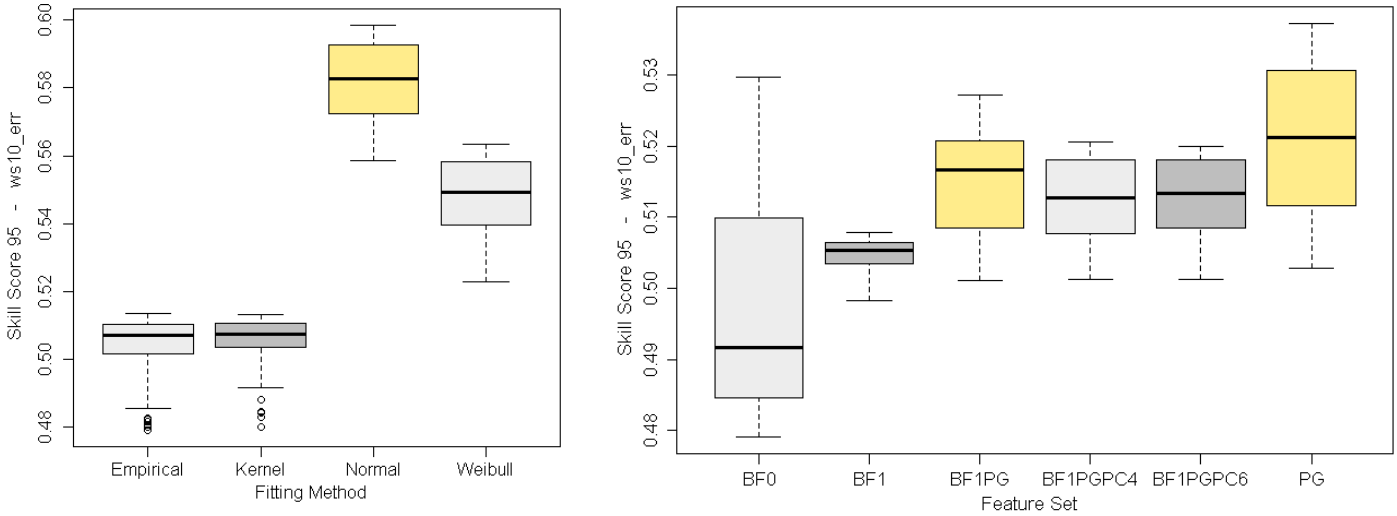


Figure 4.21. Comparison of (left) fitting methods and (right) feature sets for the wind speed PI methods for the BC data set

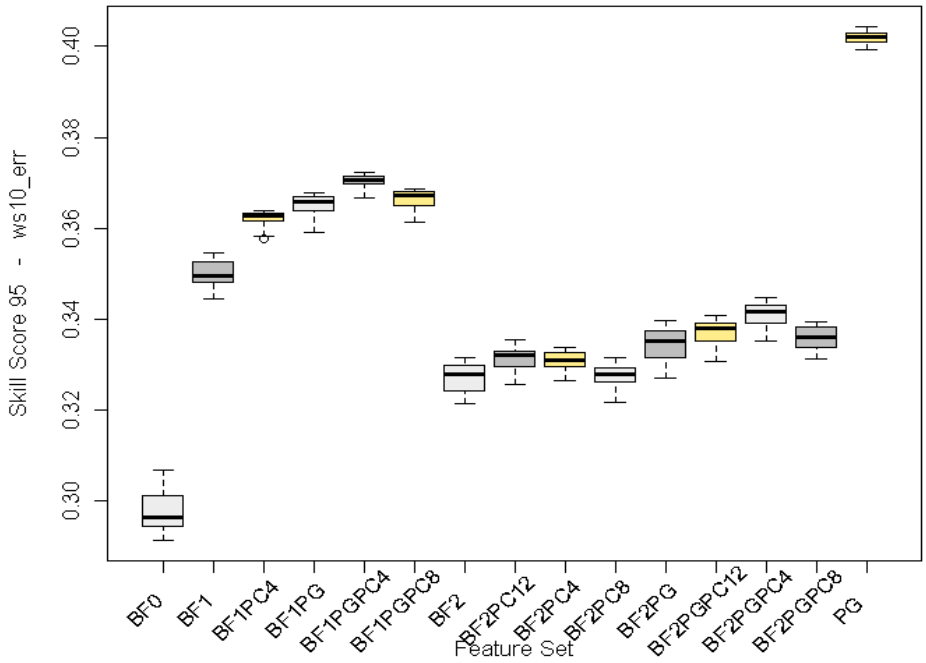


Figure 4.22. Comparison of 15 different feature sets for the wind speed PI methods for the AG data set

Table 4.8. Top five methods and the detailed measures for wind speed in AG data set based on $Sscore^{0.95}$ in 3-fold (yearly) cross validation

Algorithm	**K**	**Fit**	**Features**	**Sharpness**	**Coverage**	**Coverage$^{0.95}$**	**Resolution**	**RMSE**	**SScore**	**SScore Rank**	**SScore$^{0.95}$**	**SScore$^{0.95}$ Rank**
K-means	35	Empirical	BF0	8.99	94.73	92.83	1.91	3.03	0.2742	18	0.2915	1
K-means	25	Empirical	BF0	9.22	94.67	93.10	1.84	3.10	0.2782	28	0.2918	2
K-means	30	Empirical	BF0	9.09	94.78	93.05	1.81	3.04	0.2764	25	0.2921	3
K-means	30	Kernel	BF0	9.37	95.89	94.32	2.02	3.02	0.2772	26	0.2928	4
K-means	35	Kernel	BF0	9.29	95.99	94.29	2.11	3.00	0.2754	23	0.2928	5
Base-Ws.	10	Empirical	Ws.	10.43	95.03	94.28	0.78	3.35	0.3038	394	0.3119	51
Base-Clim.	1	Kernel	-	12.63	95.03	94.75	0.00	3.23	0.3924	4392	0.3973	2984
Base-Temp.	10	Kernel	Temp.	12.81	95.15	94.33	0.97	3.20	0.3908	4015	0.4032	3557
Base-Month	12	Kernel	Month	12.51	94.71	93.66	1.37	3.22	0.3881	3629	0.4038	3621

The detailed measures for the top five wind speed PI forecasting methods are given in Table 4.8. Here the baseline method that groups the forecast cases by wind speed bins is ranked as the 51st method among all. Yet the best skill score is achieved by K-means clustering over the newly defined simple BF0 feature set. The RMSE of wind speed forecasts would be decreased to 3.03 from 3.35 in the best baseline method after using the top clustering method.

To have a closer look at the PIs forecasted by the best temperature PI method in the BC data set, which is K-means clustering with Kernel density smoothing and *K*=6, Table 4.9 provides the details of PIs from each of the six clusters for a sample fold of test

results. As expected, the method is able to provide PIs with dynamic width. The third column of this table shows the average distance of a missed case from the edge of the forecasted PI. It is also worth to note that for cluster number one where there are fewer test cases available, the difference between the measured coverage and SScore and their respective 95%-confidence level boundaries are bigger compared to other clusters. The Kolmogrov-Smirnov goodness-of-fit test results in the last two columns also suggest that the hypothesis that trained error model and the observed test error follow the same distribution is not rejected on the 10% level for five out of six clusters.

Table 4.9. Cluster-level measures for the best temperature PI method in BC (K-means, Kernel, K=6)

Cluster No.	Width	$\bar{\delta}_M^{\alpha,j}$	Test Cases	Missed	Coverage	Coverage$^{0.95}$	SScore	SScore$^{0.9}$	K-S Test Statistic	K-S Test P-Value
1	10.98	0.73	175	9	94.86	91.20	0.3119	0.3507	0.06	0.56
2	12.11	1.85	317	14	95.58	93.18	0.3844	0.4287	0.04	0.82
3	12.68	1.44	292	12	95.89	93.43	0.3763	0.4118	0.05	0.48
4	12.71	2.13	248	9	96.37	93.75	0.3950	0.4507	0.11	0.00
5	13.33	1.13	396	24	93.94	91.58	0.4021	0.4288	0.05	0.30
6	14.85	1.18	538	21	96.10	94.43	0.4172	0.4389	0.04	0.29

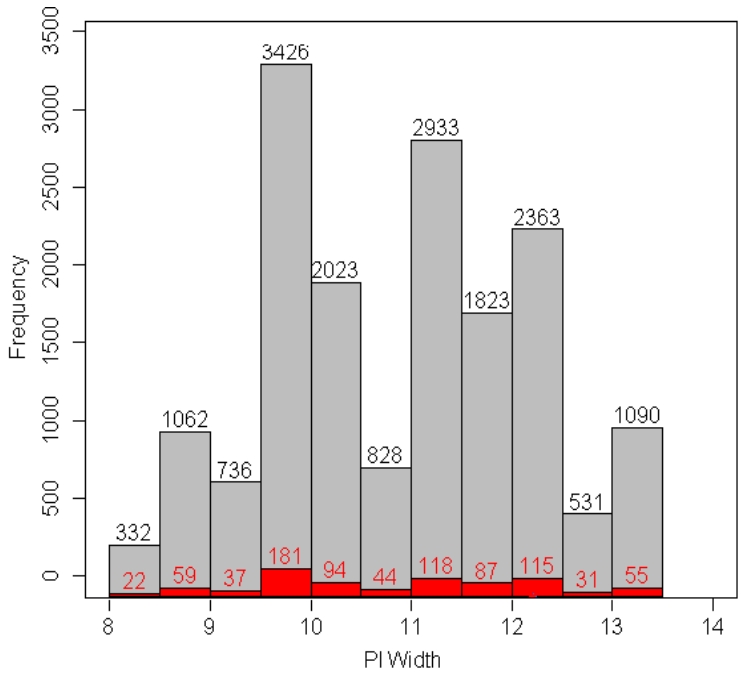


Figure 4.23. Histogram of forecasted temperature PI widths (total counts and miss cases) for the top method in AG.

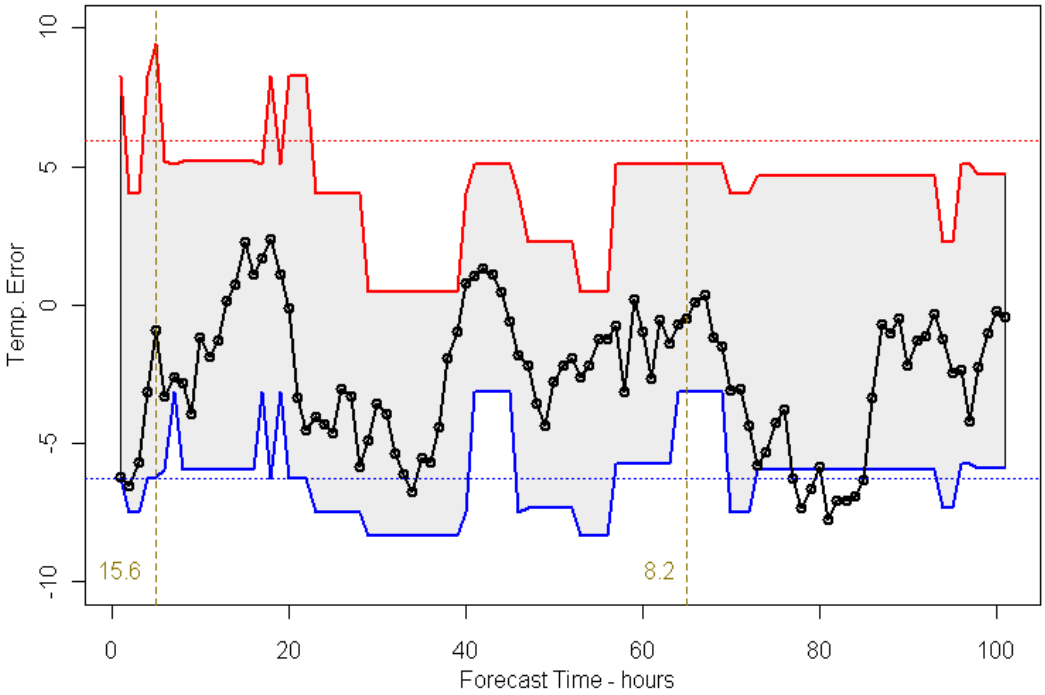


Figure 4.24. Sample temporal trends of upper (red) and lower (blue) boundaries of prediction intervals for temperature error and the actual observations (black)

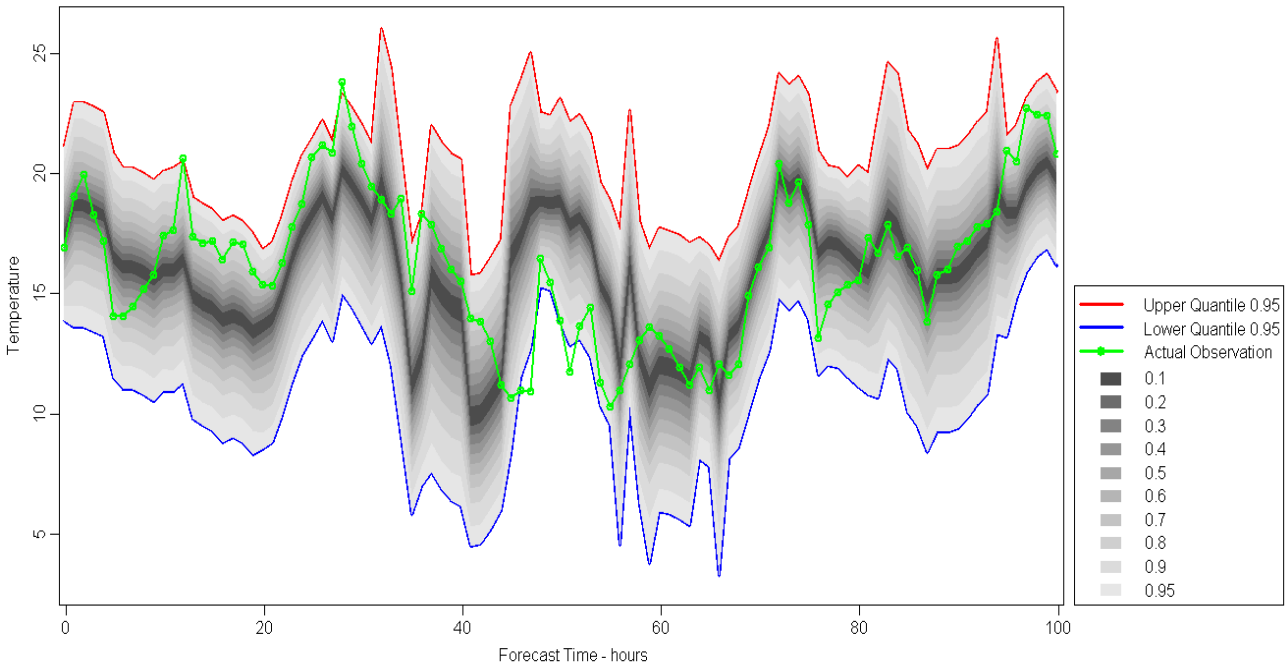


Figure 4.25. Examples of eleven different confidence level prediction intervals for temperature forecasts in 2009

For the best temperature PI forecasting method in the AG data set i.e. K-means clustering with kernel fitting and *K*=50, Figure 4.23 provide information on how variant the forecasted PIs were in terms of their width. This figure also shows the distribution of the PIs that actually missed the observed value. It is clear that the forecasted PIs have a

dynamic property and are sharper compared to the static PIs of 12.17 in the climatological baseline method. Also, it is evident that the PI forecasting system has been able to maintain a stable level of coverage in interval forecasts with different level of uncertainty (~94% in the leftmost bin and second biggest bin).

An example of PI forecasts for temperature made for 2009 by the best PI method in AG data set is depicted in Figure 4.24 for the Agassiz station over 100 consecutive hours. The horizontal lines represent the PIs of the climatological baseline method. This figure clearly shows the dynamic change of the estimated forecast uncertainty for the different predictions. PIs as narrow as 8 degrees and as wide as about 16 degrees are estimated by this method. The higher sharpness of the clustering method forecasts (i.e. smaller average PI width) is also evident in this graph. Figure 4.25 shows a fan chart of 11 different confidence level PIs (i.e. 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95) by the best method along with the observed value for temperature (not temperature error). Again the conditional nature of forecast uncertainty based on forecast situation is apparent in the forecasted PIs.

4.5.3. Fuzzy Clustering PI Forecasting Methods

In the application of fuzzy clustering both bootstrapping and distribution averaging methods of error density estimation are implemented. The obtained prediction intervals were very similar using either of these two methods. However, the latter is preferred due to a notably better efficiency. This is due to the fact that a whole process of bootstrapping and error estimation is required for every single test case in the first option. Hence, we only report the distribution averaging results here.

Figure 4.26 plots the fitted error distribution of three different fuzzy clusters obtained over the AG data set using BF2 feature set. Note that each distribution is attained by kernel density estimation using all of the samples in the AG data set along with weights from that fuzzy cluster's membership values. Using the FCM algorithm, the forecast situations are defined as fuzzy sets using the training data. As a consequence, the forecast cases are not associated with only one cluster but have different levels of membership in all clusters. The three best-performing fuzzy and non-fuzzy setups are listed in Table 4.10 and Table 4.11.

FCM was applied to the BC data set using the best-performing feature sets and fitting method for this data set based on the non-fuzzy model evaluations described in the

previous subsection (i.e. BF1, BF1PGPC4 and kernel fitting). The results show a modest improvement of PI verification score using the fuzzy approach (0.4194 vs. 0.4245).

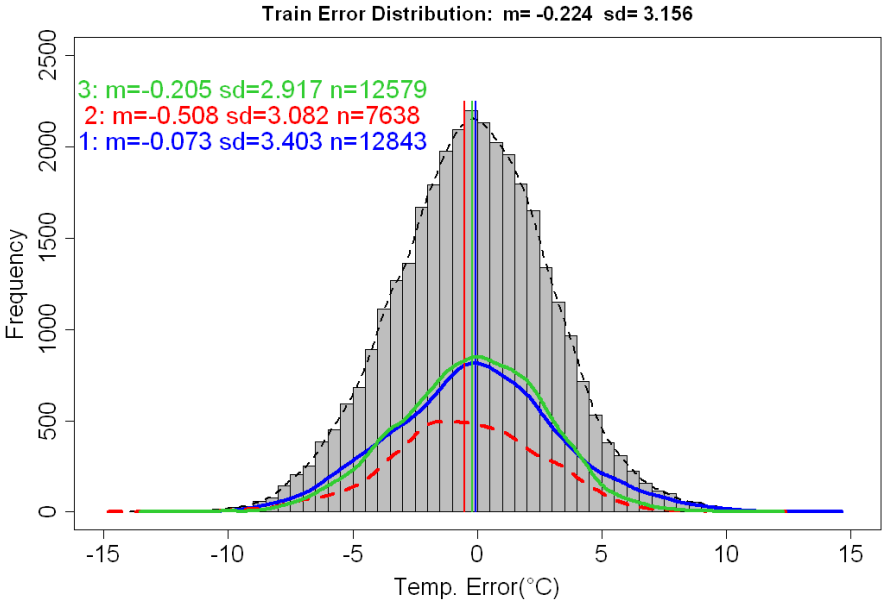


Figure 4.26. Forecast error distribution in 3 clusters of 2007 and 2008 AG data and the corresponding fitted kernel density distribution

Table 4.10. PI verification measures for top methods of temp. PI in BC data set based on 5-fold cross validation

Algorithm	***K***	**Fit**	**Features**	**Sharpness**	**Coverage**	**Coverage$^{0.95}$**	**Resolution**	**RMSE**	**SScore**	**SScore Rank**	**SScore$^{0.95}$**	**SScore$^{0.95}$ Rank**
FCM	5	Kernel	BF1	13.63	95.82	93.80	0.77	3.39	0.3934	1059	0.4194	1
FCM	8	Kernel	BF1PGPC4	13.35	95.74	93.00	0.97	3.34	0.3866	796	0.4206	2
FCM	11	Kernel	BF1PGPC4	13.12	95.70	92.35	1.04	3.29	0.3803	551	0.4213	3
K-means	6	Kernel	BF1	13.42	95.37	93.05	1.20	3.41	0.3946	1087	0.4245	13
HClust	4	Kernel	BF1	13.79	95.66	93.86	0.73	3.40	0.4014	1248	0.4251	14
K-means	9	Kernel	BF1	13.26	95.58	92.68	1.29	3.37	0.3879	865	0.4253	15
Base-Temp.	10	Kernel	Temp.	14.23	95.59	92.59	1.10	3.65	0.4071	1368	0.4423	489
Base-Clim.	1	Kernel	-	14.99	95.19	94.32	0.00	3.80	0.4394	1955	0.4514	829
Base-Ws.	10	Kernel	Ws.	14.49	95.61	92.89	0.92	3.67	0.4235	1487	0.4621	1001
Base-Hour	24	Kernel	Hour	13.72	95.77	90.21	1.09	3.47	0.3972	1111	0.4679	1098

A similar skill improvement was observed for the AG data set using BF2 and BF2PG feature sets and the kernel density smoothing method. PIs obtained by the FCM algorithm (with K=45) have the best skill score and the least RMSE when considering point forecasts. By transforming the periodic variables of wind direction and hour of day as explained in subsection 4.5.1, the FCM can further improve SScore and SScore$^{0.95}$ into values of 0.3173 and 0.3401, respectively. The FCM-based PIs have rather smaller values

of resolution. This is expected as in these models the error characteristics of every forecast case are affected by *all* discovered situations although with different intensities. However, the presented empirical evaluation study confirms that the proposed clustering-based methods improve the skill of forecasted PIs compared to baseline methods.

Table 4.11. PI verification measures for top methods of temp. PI in AG data set based on 3-fold (yearly) cross validation

Algorithm	*K*	Fit	Features	Sharpness	Coverage	Coverage$^{0.95}$	Resolution	RMSE	SScore	SScore Rank	SScore$^{0.95}$	SScore$^{0.95}$ Rank
FCM	45	Kernel	BF2	10.62	94.89	92.77	1.59	2.77	0.3220	27	0.3432	1
FCM	30	Kernel	BF2PG	10.91	94.93	93.26	1.65	2.86	0.3285	106	0.3452	2
FCM	50	Kernel	BF2PG	10.67	94.78	92.49	1.79	2.81	0.3231	41	0.3459	3
K-means	50	Kernel	BF2	10.78	94.96	92.74	1.87	2.80	0.3254	64	0.3485	13
K-means	45	Kernel	BF2	10.86	94.89	92.78	1.87	2.83	0.3273	87	0.3492	15
K-means	40	Kernel	BF2	10.89	94.82	92.85	1.84	2.83	0.3303	128	0.3499	16
Base-Month	12	Kernel	Month	12.21	95.12	94.10	1.91	3.12	0.3601	2588	0.3704	1719
Base-Temp.	10	Normal	Temp.	11.70	94.44	93.57	0.98	3.04	0.3620	2809	0.3725	2193
Base-Ws.	10	Kernel	Ws.	12.12	94.91	94.17	1.20	3.12	0.3664	3037	0.3754	2681
Base-Clim.	1	Normal	-	12.17	94.78	94.49	0.00	3.11	0.3740	3985	0.3774	2899

4.6. Conclusions

Forecast uncertainty plays an important role in many practical applications of meteorology. In this study, the historical performance of WRF NWP model is used as a source of information for uncertainty modeling. The proposed approach allows dynamic analysis of uncertainty based on context, i.e., predicted weather situation. Contexts of weather forecasts are established by automatically discovered clusters and then used to derive conditional PIs through statistical analysis. The effectiveness of the proposed approach has been empirically evaluated using two data sets of weather hindcast and associated observations.

Several feature sets were applied to group weather situations using four different clustering algorithms (K-means, Clara, HClust and Fuzzy C-means). To assess the proposed PI computation methods, we created a comprehensive evaluation framework based on a proper skill score metric. The assessment results confirmed the applicability of the proposed PI computation methods and showed that the resulting PIs have high sharpness and skill.

Comparisons to various baseline methods confirm an average 8% improvement in PI forecast skill when using the proposed dynamic methods based on Fuzzy C-means clustering. As a result of their nature, the proposed methods also intrinsically remove bias, decreasing the RMSE of point forecasts by up to 10%. The proposed PI modeling

methods can be used in real world applications to enhance point forecasts of NWP systems with information on prediction uncertainty.

Chapter 5
Quantile Regression Approaches to Uncertainty Modeling

In this chapter a range of quantile regression methods are investigated as tools for learning forecast uncertainty models. Specifically, for the first time a hybrid approach of clustering and kernel quantile regression is applied in the context of weather forecast uncertainty modeling. The prediction intervals obtained from all of the quantile regression models are practically examined and then compared to the clustering methods using a real data set of NWP forecasts.

5.1. Introduction

The degree of uncertainty in the forecasts of a Numerical Weather Prediction (NWP) model can potentially have an enormous impact on the decisions that are made based on these forecasts. Wind power production and marketing [66], Dynamic Thermal Rating (DTR) systems used by power transmission utilities [34][71], and extreme weather event prediction systems [70] are just few example applications where the forecast uncertainty is often regarded as significant as the expected forecast value itself [37][53].

NWP models are advanced computer simulation systems that provide expected values of various weather attributes, on a three dimensional spatial grid and at certain forecast horizon [80]. These systems do not provide any information about the uncertainty of the forecasts. However, there is always some level of error associated with forecasts and the degree of this inaccuracy is known to be variable for different predictions [48]. Imprecision of initial conditions, parameterization of sub-grid scale processes, and various simplifying assumptions incorporated in the NWP system are regarded as some major reasons for forecast inaccuracies [61].

Although the raw outputs of an NWP system provided as point predictions can be easily understood and evaluated, the probabilistic nature of the forecasts (which can represent the prediction uncertainty) is dismissed. Prediction Intervals (PI) are a prominent form of forecast uncertainty communication. They are defined as a value interval accompanied by a confidence level for actual observations to be inside this interval (e.g. T = [-3°C, 10°C], conf = 95%) [16][29][68].

There is a large body of literature on obtaining such uncertainty information from NWP models using ensemble forecasting systems [20][72][60][81]. However, ensemble predictions may incur large computational costs, making them infeasible in some cases. Additionally, instant availability of historical performance data sets for many existing forecasting systems and potentially useful uncertainty patterns hidden in them has made post-processing approaches to uncertainty modeling an increasingly attractive topic [3][7] [59][69] [68].

Error distribution fitting and clustering methods have been studied recently as major methods towards learning forecast uncertainty models from historical system accuracy data sets [68]. These methods rely on the known fact that different forecast situations typically exhibit different levels of forecast uncertainty, and that such patterns can be potentially found from the system performance record [48]. Hence, the forecasts of the system are first clustered into similar groups using related attributes regarded as influential variables. Next, the historical error distribution of each cluster is modeled by fitting either a parametric (e.g. Gaussian) or a non-parametric (e.g. empirical) distribution. The desired quantiles of a new forecast are then calculated from the fitted error distribution of the cluster which the new forecast case belongs to. Hence, these approaches can provide uncertainty information for a point forecast in the form of a full probability distribution. The distribution can then be used to obtain prediction intervals with any desired level of confidence.

In quantile regression based methods, on the other hand, each individual quantile is modeled independently and there are no assumptions on the distribution of the forecast error [59][90]. These methods can learn a direct relationship between the target quantile and the set of available influential attributes through an optimization process. Various quantile regression methods have been proposed and applied to forecast uncertainty modeling. Bremnes [6] proposed the application of local quantile regression to obtain non-linear models of quantiles for wind power forecasts. In another study, Nielsen *et al.*

[59] applied an additive quantile model by using spline basis functions. In both works, the resulting prediction intervals were evaluated in terms of their inter-quantile range and actual observation frequencies, as compared to the forecasted quantile. Yet, the skill of the prediction interval forecasting system was not evaluated in an objective framework. A few statistical models including local quantile regression were compared in a study by Bremnes [7] as quantile forecasting models for wind power from NWP outputs. Sharpness and reliability measures were evaluated for different setups of these models but forecast skills were not verified. Pinson and Kariniotakis [68] demonstrated a novel fuzzy inference model based on grouping of forecasts and adapted resampling for distribution fit. A detailed comparative study of this method and one quantile regression-based method was provided in Pinson *et al.* [69]. An improved version of this approach using fuzzy clustering and error distribution fitting was introduced in Chapter 4. Additionally in the domain of statistical methods, time-adaptive kernel density estimation methods were proposed by Bessa *et al.* [3]. In this research, we aim for a comprehensive comparative study for the application of some major quantile regression methods and recently proposed distribution fitting methods. In addition, the relatively new kernel quantile regression method [51] is also investigated in the context of weather forecast prediction interval modeling in this study. Due to performance issues a hybrid clustering-quantile regression approach is proposed for application of this method. A comparative study is performed using a large, real-world NWP data set with a focus on forecast skill as a significant measure for essential conclusive comparisons between prediction interval forecasting systems. As discussed previously, skill score measurements are extended by considering sampling uncertainties in the test experiments as explained in Chapter 3. This approach also offers a good foundation to investigate the role of different parameters involved in these methods.

5.2. Linear and Non-linear Quantile Regression for PI Modeling

In linear regression tasks, a target variable is estimated by a linear combination of a set of related features. The unknown coefficients of this linear equation are tuned by an optimization process using an objective function (e.g. squared error in real value regression). The same approach can be used to find a linear relationship between a set of features and a specific "quantile" of a target variable. The θ-quantile of the target variable y denoted as $\hat{q}_y^{\theta}$ is formulated as [44][45] [54]:

$$\hat{q}_y^{\theta} = f(\mathrm{x}) = \beta_0^{\alpha} + \beta_1^{\theta} x^1 + \beta_2^{\theta} x^2 + \cdots + \beta_d^{\theta} x^d \quad (5.1)$$

where $x^j, j = 1..d$ are the d influential variables for modeling the θ-quantile of y, and $\beta_j^{\theta}, j = 1..d$ constitutes the β_y^{θ} vector of coefficients for the target θ-quantile. This vector is estimated using the following optimization objective [44][90]:

$$\hat{\beta}_y^{\theta} = argmin_{\beta} \sum_{i=1}^{N} L_{\theta}\left(y_i - \left(\beta_0^{\theta} + \beta_1^{\theta} x_i^1 + \cdots + \beta_d^{\theta} x_i^d\right)\right) \quad (5.2)$$

where $i = 1..N$ is the number of recorded pairs (y_i, x_i) in the data set, and L_{θ} is the loss function of a θ-quantile target defined as:

$$L_{\theta}(\delta_i) = \begin{cases} \theta\delta_i & \delta_i \geq 0 \\ (\theta - 1)\delta_i & \delta_i < 0 \end{cases} \quad and \quad (5.3)$$

$$\delta_i = y_i - \hat{q}_{y_i}^{\theta} \quad (5.4)$$

The optimization task formulated in Equation (5.2) is then solved using linear programming techniques [44][54]. In order to obtain $\hat{I}_{new}^{\alpha}$, quantiles of target error, $\hat{q}_e^{\theta_l}$ and $\hat{q}_e^{\theta_u}$ are separately modeled by linear quantile regression using the data set of (e_i, x_i), where e_i represents the recorded error of forecast case i, and x_i is the vector of explanatory influential variables. This yields the optimizer vectors of $\hat{\beta}_e^{\theta_l}$ and $\hat{\beta}_e^{\theta_u}$ that are then used to compute the $(1-\alpha)$-confidence level prediction interval of target y for any new forecast x_{new}:

$$\hat{q}_{y_i}^{\theta_l} = \langle \hat{\beta}_e^{\theta_l}, \mathrm{x}_{new} \rangle + \hat{y}_i, \;\; \hat{q}_{y_i}^{\theta_u} = \langle \hat{\beta}_e^{\theta_u}, \mathrm{x}_{new} \rangle + \hat{y}_i \quad (5.5)$$

where $\langle .,. \rangle$ represents the dot product of the two vectors. Note that an entry of 1 should be added as the leftmost element of X_{new} to be multiplied by the $\hat{\beta}_0^{\alpha}$ term as the intercept. As opposed to the methods described in Chapter 4, there is no distribution fitting process required in further steps when using quantile regression. This also means that a new model has to be trained for any new quantile of interest.

The model optimized and stored in $\hat{\beta}_y^{\theta}$ describes a linear relationship between the error quantile and the influential features in x. However, using a non-linear transformation basis function $\phi(\mathrm{x})$ to derive new features from the currently available features, one can in effect learn non-linear relationships by still using the linear formulation in Equation (5.2). For instance, to learn n^{th} degree polynomial functions in quantile models, one can extend features in x by adding $\phi(\mathrm{x}) = [\mathrm{x}^2, \mathrm{x}^3, \ldots, \mathrm{x}^n]$ and then perform the same process of optimizing the linear relationship between the new feature

set and the target quantile. More complex forms of non-linear relationships can also be represented and optimized using other basis functions such as the sigmoid and $\sin(\mathrm{x})$ etc. In the experiments, we use the second and third degree polynomials along with the sine. We use NLQR to refer to the prediction interval modeling methods that use these transformed features, while LQR refers to the methods that use the original vector x.

5.3. Quantile Regression with Spline-basis Functions

Additive quantile models are another technique used by Nielsen *et al.* [59] to learn non-linear models of weather forecast quantiles. Spline functions are the most frequently used basis functions [3][54]. Since it is expected that the relationships between forecast values and forecast errors are of a non-linear nature, spline-basis can provide a suitable transformation.

This relationship can then be approximated by a linear combination of basis functions of the influential attributes [32]:

$$\hat{q}_y^{\theta} = \beta_0^{\theta} + \sum_{j=1}^{d} \sum_{k=1}^{df_j} \beta_{j,k}^{\theta} f_{j,k}(x_j) \quad (5.6)$$

where $f_{j,k}$ is the cubic B-spline basis function used for feature j using df_j degrees of freedom. Note that assuming a constant *df*-degree of freedom for all basis functions, there will be $df \times d$ features in the final model to be optimized by the linear optimization task formulated in Equation (5.2). The appropriate value for this parameter has to be determined based on experiments involving the training data set.

5.4. Local Quantile Regression

Unlike spline additive models of quantile regression, in local quantile regression (LocQR) there is no effort made to learn complex non-linear models for a quantile. Instead, it is assumed that, in the close neighborhood of a given x, the relationship between x and the target quantile is simple enough to be modeled linearly. Rationally, data points that are closer to x should have more impact on this linear model than those further away from it. This can be formulated as the following optimization problem [91]:

$$\hat{\beta}_{e,\mathrm{x}}^{\theta} = argmin_{\beta} \sum_{i=1}^{N} L_{\theta}\left(y_i - \beta^{\theta}(\mathrm{x_i} - \mathrm{x})\right) w(\mathrm{x_i}, \mathrm{x}) \quad (5.7)$$

where $\hat{\beta}_{e,\mathrm{x}}^{\theta}$ is determined for input x by considering a set of training samples that are centered around x and weighted using [6]:

$$w(\mathrm{x_i},\mathrm{x}) = \begin{cases} \left(1-\left(\frac{\|\mathrm{x}_i-\mathrm{x}\|}{d_\lambda(\mathrm{x})}\right)^3\right)^3 & \text{if } 0 \le \frac{\|\mathrm{x}_i-\mathrm{x}\|}{d_\lambda(\mathrm{x})} < 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.8)$$

where $d_\lambda(\mathrm{x})$ is the distance from x to its λN-th nearest neighbor among training samples $\mathrm{x}_{1..N}$ and $\|\mathrm{x}_i - \mathrm{x}\|$ is the Euclidian distance between the two vectors. Based on this definition, $(1-\lambda)N$ data samples have zero weight and hence have no impact on the quantiles optimized at point x.

Note that using LocQR, two new models have to be optimized for *each* new forecast (x) to compute the upper and lower quantiles of the prediction interval for that specific forecast. This is in contrast with the scenarios of applying other regression methods described above, in which these models are learned only once and then utilized to provide prediction intervals for *any* future forecasts.

5.5. Kernel Quantile Regression

To learn arbitrarily complex nonlinear models, the optimization process can be performed in reproducing kernel Hilbert spaces (RKHS) leading to kernel quantile regression (KQR) [51][76]:

$$\hat{\beta}_y^\theta = argmin_\beta \; C\sum_{i=1}^N L_\theta\left(y_i - \beta^\theta x\right) + \frac{1}{2}\|\beta\|^2 \quad (5.9)$$

where the last term is obtained from the RKHS norm of function g ($\|.\|_{\mathcal{H}}^2$) and f=g+β_0^θ, where function g only contains $\beta_{1..d}^\theta$ so that here the constant offset is not regularized in the above objective. The regularizer penalizes more complex functions to avoid overfitting. C is the cost factor that balances the total loss over this penalization. The above formulation is very similar to the well-known primal form of support vector regression [83], except for the loss function that has been redefined to optimize for the conditional quantile of interest rather than the conditional mean or median.

This formulation also allows obtaining a dual form of the optimization problem using Lagrange multipliers that would represent the model by vector of weights ($\alpha_i, i = 1..N$) over samples (rather than features in the primal problem) [77]. Since the dual form only uses the vector products of the input vectors, we only need to consider the kernel function (k) which would provide an inherent Φ-mapping of inputs into a new feature space:

$$\hat{\alpha}_y^\theta = argmin_\alpha \frac{1}{2}\alpha^T K\alpha - \alpha^T\vec{y} \;\; subject\ to\ C(\theta-1) \le \alpha_i \le C\theta \;\; for\ all \;\; 1 \le i \le N \;\; and \;\; \vec{1}^T\alpha = 0 \quad (5.10)$$

$$K_{ij} = k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \quad (5.11)$$

where the kernel matrix K is positive-semidefinite. Note that $\Phi(x_i)$ does not appear explicitly in the objective function and only the inner products of the transformed vectors (represented as K_{ij} entries in the kernel matrix) are used. The above dual form can be solved using quadratic programming and the f function can then be recovered. A common choice for the kernel function is the Gaussian kernel [77]:

$$k(x_i, x_j) = exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right) \quad (5.12)$$

where $\sigma > 0$ is the kernel width parameter and should be tuned.

Because of the low scalability of this method (the kernel matrix is N by N), we propose a two-step process where the training data is first clustered into K partitions using the same feature set and kernel quantile regression is applied independently to each partition. In the test phase, a new forecast is first assigned to its closest cluster and then passed to the learned model for that cluster to obtain the quantiles.

5.6. Experimental Results

5.6.1. Data Sets and Method Set-ups

The data set used in this set of experiments is the AG data set described in subsection 4.5. Also the utilized feature sets (other than new feature sets defined in this section) are defined in subsection 4.5.

Table 5.1. feature set definition in PI models using combinations of basic features

Feat Set	m	d	h	t2	ws	wd	sp	pg
C1				●	●			
C2				●	●		●	
C3			●	●	●		●	
C4		●		●	●		●	
C5	●			●	●		●	
C6	●	●	●	●	●		●	
C7	●	●	●	●	●	●	●	●

Table 5.1 describes seven new combinations from the basic features and pressure tendency. These feature sets are defined to perform a more detailed analysis on the role of basic attributes, specifically month, day and hour. Three-fold cross validation was used by splitting different years into folds. For instance, the 2007 and 2008 data was used to train the prediction interval model and then prediction intervals for 2009 obtained by this

model were evaluated for their quality and skill. Random-based 5-fold cross validation experiments were also conducted. However, due to the similarity of the obtained results, only year-based cross validation results are reported here.

5.6.2. Forecast Evaluation Results

In quantile regression models, the forecasts are clustered using different numbers of clusters in the range from 2 to 100 for projection of the skill score sampling variance analysis and confidence bound computation. The role of number of degrees of freedom in spline-basis as an important parameter of the spline quantile regression (SPQR) models is depicted in Figure 5.1. The curves show the change of $SScore\ _{M}^{0.95}$ using K=50 clusters for models using different feature sets. The number of degrees of freedom by which a model achieves its best score is encircled. Note that K=50 is chosen as it was the number of clusters which best represented the forecast groups in experiments involving clustering-based methods.

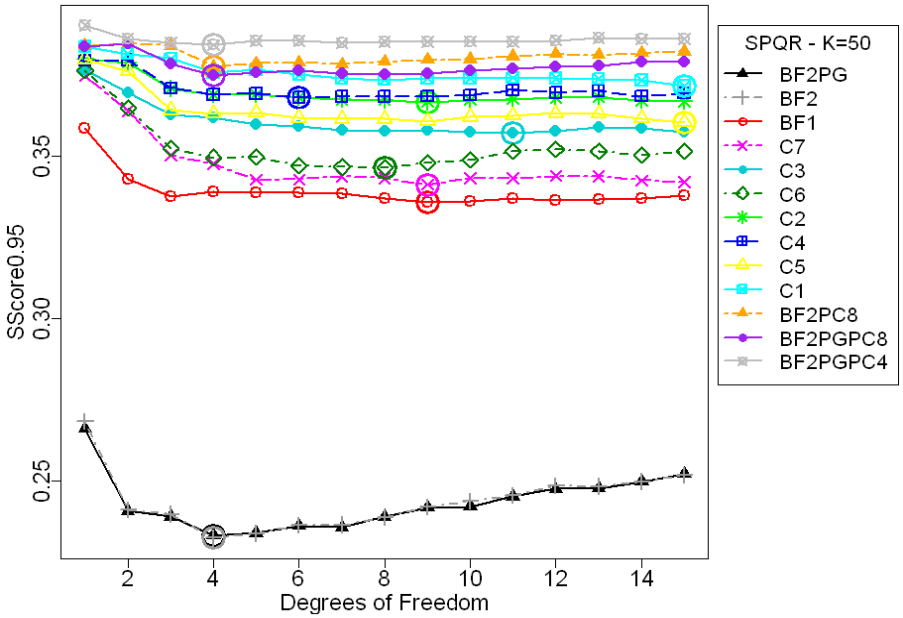


Figure 5.1. Projection of $SScore\ _{M}^{0.95}$ for spline quantile regression models over different degrees of freedom using various feature sets

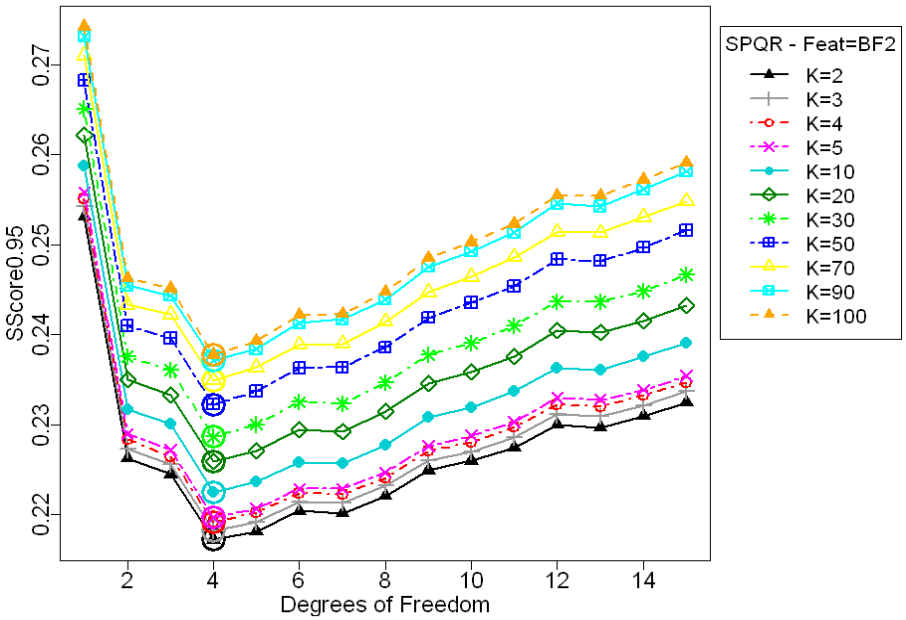


Figure 5.2. Projection of $SScore\ _{M}^{0.95}$ for spline quantile regression models over different degrees of freedom using various number of clusters used in skill score uncertainty analysis

BF2 and BFPG feature sets provide the best models using the SPQR method by a considerable margin compared to the other feature sets. These models achieve the best skill using 4 degrees of freedom. In Figure 5.2 different curves show the trends of skill score over degrees of freedom when using different number of clusters in skill score uncertainty analysis. This figure shows that the SPQR model with four degrees of freedom achieves the best score for all alternative numbers of clusters used in sampling analysis.

In local quantile regression (LocQR) models the role of λ is investigated. As previously mentioned, there is essentially no offline training phase involved in LocQR and two quantile models have to be optimized for every single test case. Experiments revealed that a rather long computational time is required for the evaluation of the whole test data set due to these characteristics. As an alternative approach, the LocQR model was trained for a limited number of points (knots - rather than every test point) randomly selected from the training samples. In the test phase, rather than training a new model for every test case, the model already trained for the nearest knot to the current test case is applied to compute the prediction interval for that test case. Different numbers of knots: 10, 100, 1000, 3000 and 20000 (indicating original LocQR with no knot selection used) were examined. The skill score of prediction interval models using BF2 and C3 feature sets are plotted in Figure 5.3 and Figure 5.4. These figures show that for BF2 feature space, that has a higher dimensionality, a larger neighborhood ($\lambda = 0.7$) is preferred by

the LocQR model. In contrast, a smaller neighborhood ($\lambda = 0.1$) is preferred for the lower dimensionality feature space of C3. Results of this set of experiments also confirm that by using a limited number of knots (e.g. 3000), the training and evaluation phase can be performed much faster without significantly compromising the accuracy of the model.

For KQR algorithm, the Gaussian kernel was chosen as the best kernel function. Tuning of this method is performed using a grid search over the two parameters σ and C. Figure 5.5 shows two curves as a sample of the grid search results that project the skill of prediction intervals over one of these parameters, while keeping the other constant. Note that due to the very large size of the Kernel matrix ($N \times N$) used in KQR, Cholesky decomposition is applied to compress this matrix into a lower rank matrix which has a feasibly computable size [23].

To explore the impact of different feature sets on the trained prediction interval models, Figure 5.6 summarizes the distribution of skill scores obtained by different models using these predictor sets. The BF2 feature set clearly provides better skill prediction interval models on average (lower $SScore^{0.95}$). The two feature sets of BF2 and BF2PG include the horizontal and vertical wind speed elements at five pressure levels. A possible explanation for obtaining better uncertainty models using these feature sets is the availability of relevant information describing the instability of the forecast atmospheric situation. Among the newly defined basic combinations, C3 attains the best score emphasizing the significance of temperature, wind speed and hour-of-day attributes.

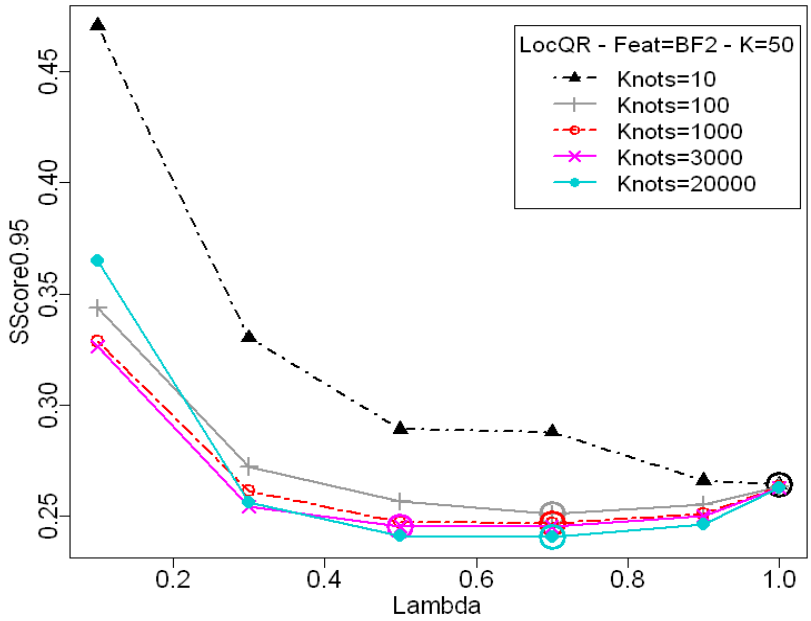


Figure 5.3 Skill score diagrams of LocQR models as a function of lambda and number of knots for BF2 feature set

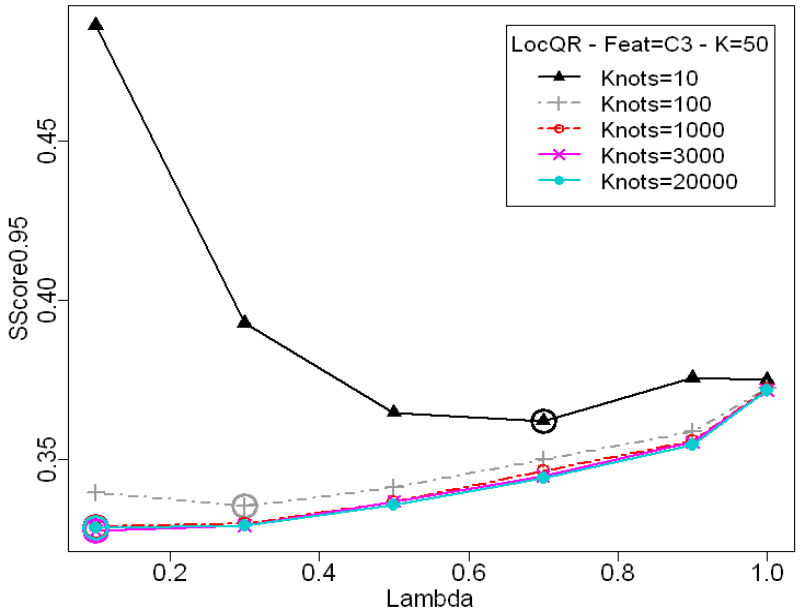


Figure 5.4. Skill score diagrams of LocQR models as a function of lambda and number of knots for C3 feature set

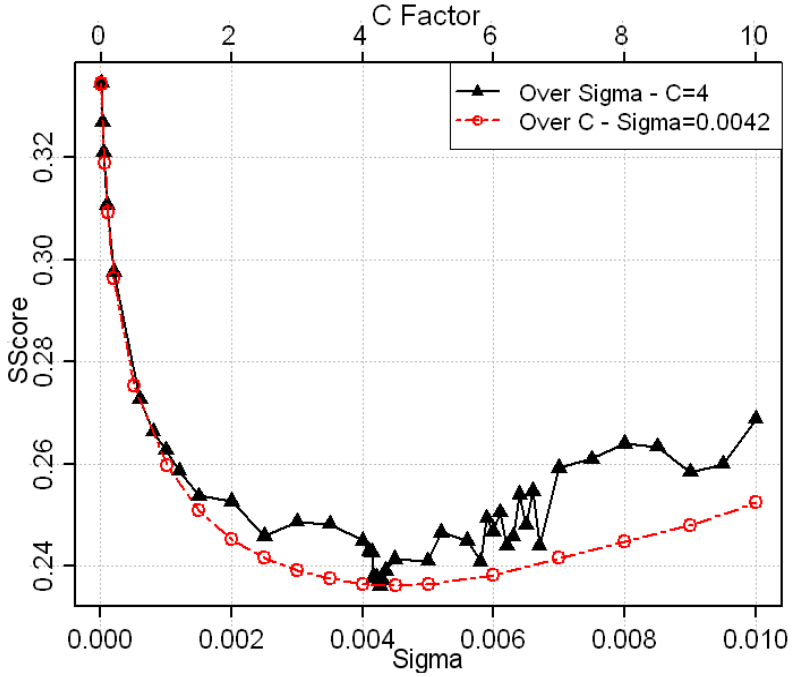


Figure 5.5. Tuning the sigma parameter in the KQR kernel function

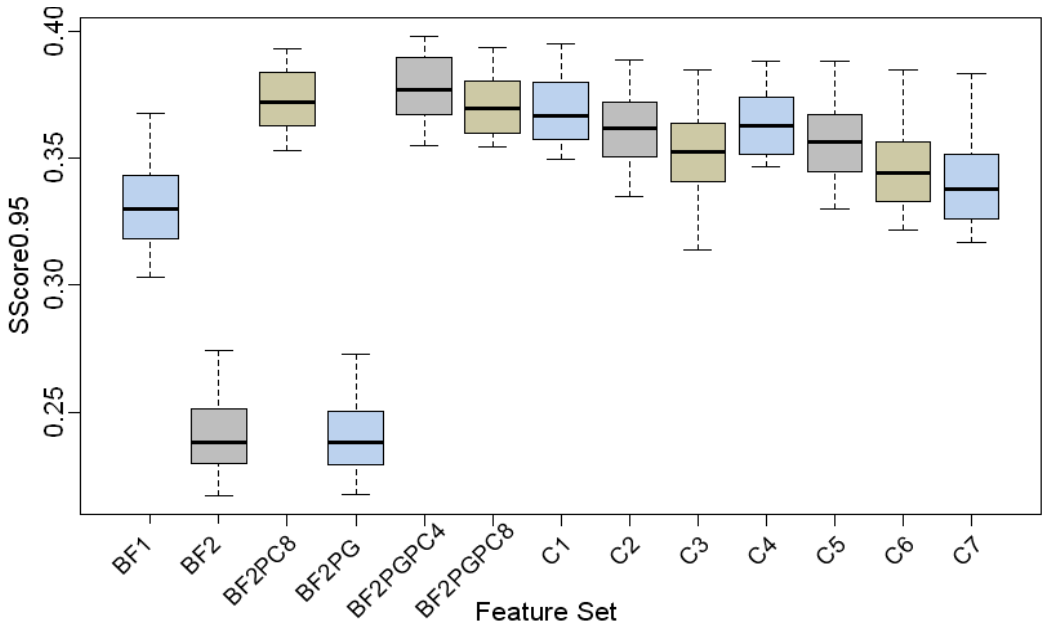


Figure 5.6. Box plot of skill score for different feature sets used by the various quantile regression methods

The details of prediction interval quality measures from (yearly) 3-fold cross validation of different methods are reported in Table 5.2. The first four columns determine the best model among each of the different quantile regression, clustering-based and baseline methods. Basic quality measures are reported in the next five columns. The 95% confidence lower bound of the coverage measure is also calculated using one-sided Binomial test. Thus a cluster with 90% coverage (hit rate) in 1000 test cases has a bigger lower bound (i.e. $Coverage^{0.95}$=88.3%) as compared to a cluster with the same 90% coverage, but with only 200 test cases (i.e. $Coverage^{0.95}$=85.8%). Moreover, Root Mean Squared Error (RMSE) is reported as a key measure for point based forecast evaluations. Note that the median of each prediction interval is considered as the new point forecast of the trained model. It should also be noted that since the upper and lower quantile models are learned independently in quantile regression approaches, they may cross one another in some cases. Although there were only few such cases (e.g. about 61 for 20% confidence level in NLQR), they were substituted by the climatological baseline prediction interval to keep a balanced judgment between different models.

The best prediction interval model is SPQR with four degrees of freedom using BF2 feature set. It is followed by LocQR, NLQR, KQR and LQR, respectively. All of the quantile regression models outperform the best fuzzy clustering based method with 45 clusters and kernel density smoothing (in terms of $SScore_M^{0.95}$). Yet, all of these learning-based models surpass the baseline methods ($p<0.0005$). The quantile regression models

provide significantly sharper prediction interval forecasts (lower average prediction interval width) but with about 3% less reliability. However, refereeing to skill score that summarizes the overall performance of a model, one can conclude on the higher quality of prediction interval forecasts by these models. By using transformed features for the cyclic attributes of wind direction and hour of day in SPQR, the SScore and $SScore^{0.95}$ measures improve to the values of 0.2119 and 0.2314, respectively. Figure 5.7 takes a detailed look at the width of forecasted prediction intervals by the different models. The constant width of climatological baseline model is shown as the horizontal line. This figure shows the sharpness of forecasts provided by quantile regression models, and specifically the SPQR model.

A fan chart showing SPQR (*df*=4) temperature prediction intervals with a range of confidence levels for a specific time frame and station is provided in Figure 5.8. One can notice the dynamic change of estimated forecast uncertainty depending on the various forecast situations.

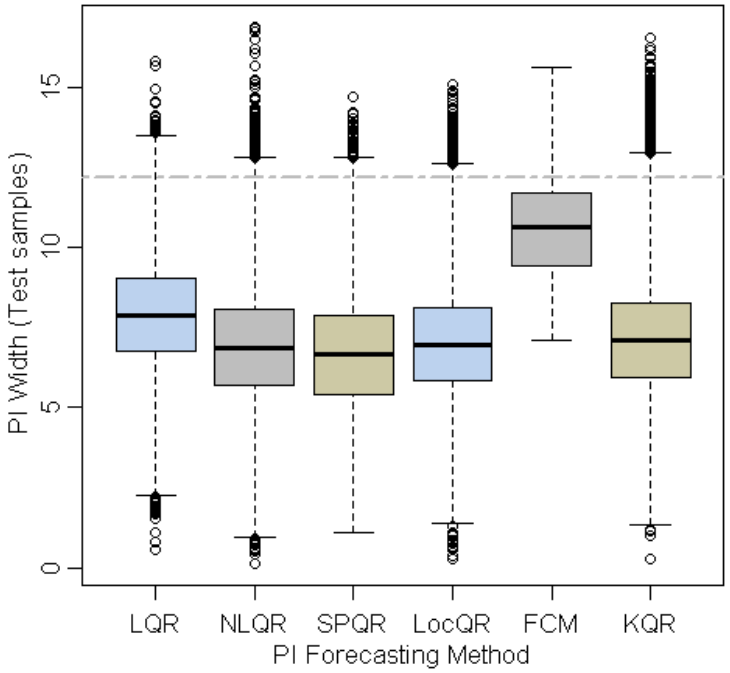


Figure 5.7. Empirical width distribution of forecasted 95% prediction intervals (horizontal line shows the best baseline model)

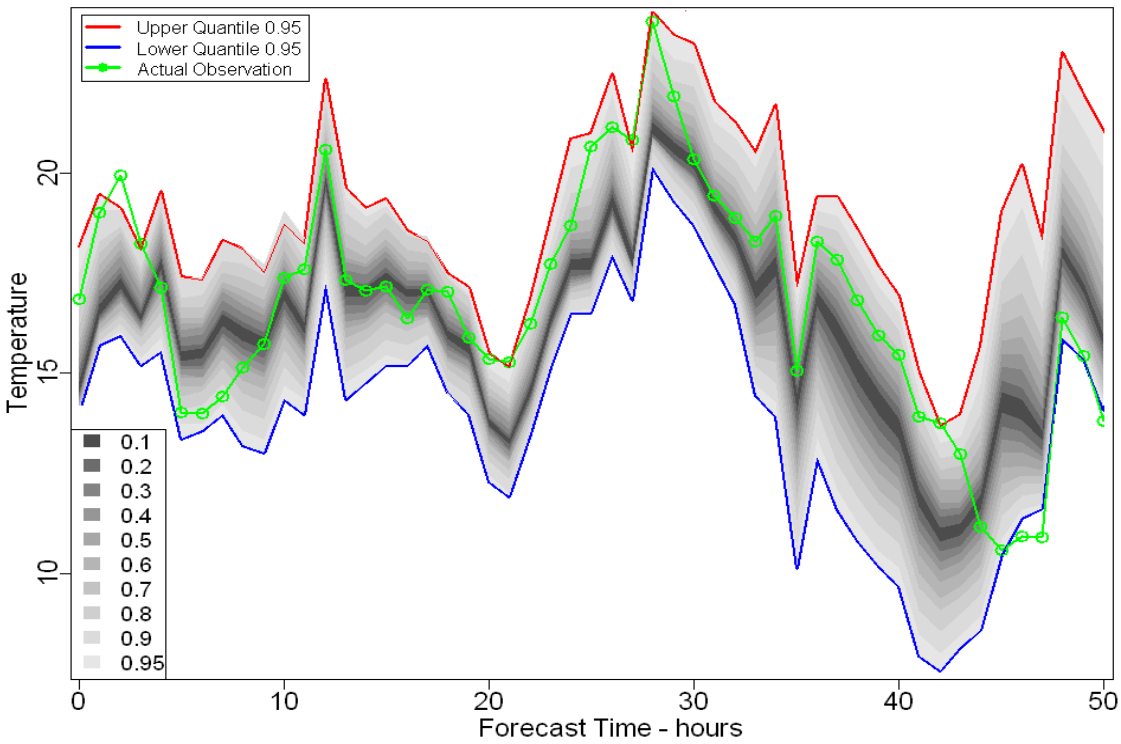


Figure 5.8. Trends of various confidence level prediction intervals and the actual observations

Table 5.2. Prediction interval verification measures for top models of different methods based on 3-fold (yearly) cross validation

Algorithm	K	Features	Fit/ Params	Sharpness (°C)	Coverage %	Coverage$^{0.95}$ %	Resolution	RMSE	*SScore*	*SScore*$^{0.95}$
SPQR	(50)	BF2	df=4	6.68	93.56	91.10	1.76	1.92	0.2125	0.2323
LocQR	(50)	BF2	λ=0.7	6.92	93.46	90.97	1.73	2.00	0.2202	0.2406
NLQR	(50)	BF2	-	6.92	93.15	90.62	1.79	2.00	0.2264	0.2492
KQR	(50)	BF2	σ=0.0042 C=4	7.16	93.09	91.51	1.85	2.05	0.2362	0.2561
LQR	(50)	BF2PG	-	7.91	94.39	92.05	1.64	2.17	0.2438	0.2640
FCM	45	BF2	Kernel	10.62	94.89	92.77	1.59	2.77	0.3220	0.3432
Base-Month	12	Month	Kernel	12.21	95.12	94.10	1.91	3.12	0.3601	0.3704
Base-Temp.	10	Normal	Temp.	11.70	94.44	93.57	0.98	3.04	0.3620	0.3725
Base-Clim.	1	-	Normal	12.17	94.78	94.49	0.00	3.11	0.3740	0.3774

Finally, Figure 5.9 depicts curves of $SScore_{M}^{0.95}$ for increasing number of clusters. This figure also confirms that the differences between skills of the best-performing models are not due to chance and that the prediction intervals obtained by SPQR are truly superior to other models. As a counter example, this is not the case between LocQR models with $\lambda = 0.5$ and 0.7. Although the first model has a better skill score in the ordinary test results, its skill score confidence bound is increasingly worse than the

second model when taking sampling variations into account within increasing number of clusters.

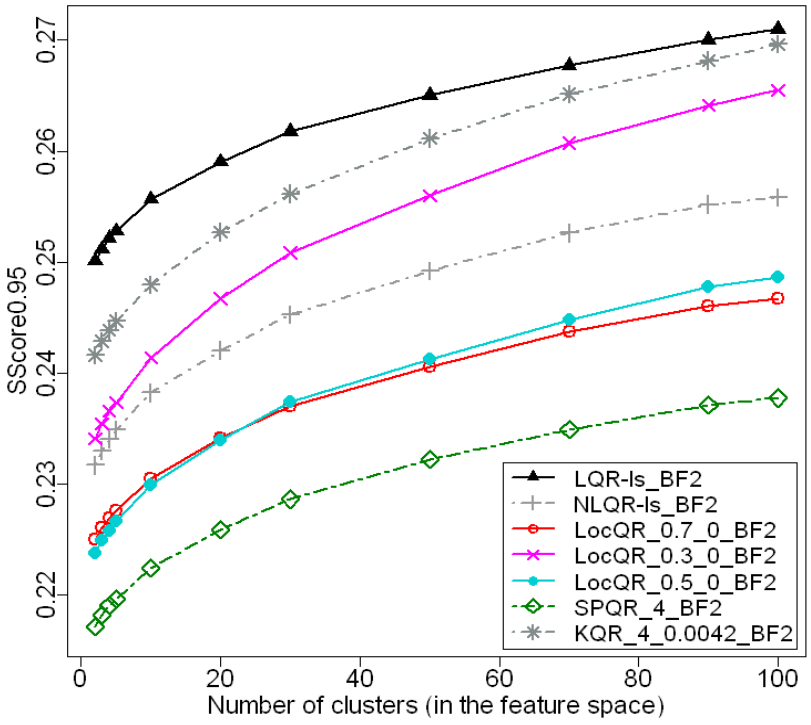


Figure 5.9. Trends of $SScore\ _{M}^{0.95}$ for the top quantile regression models

In other words, the better skill of the first model in the initial test results is most probably merely due to chance (by providing good prediction intervals in the areas that insufficient samples are available to evaluate the model). This example signifies the role of skill score uncertainty analysis for real-world evaluations and decisions.

It is also important to note that by modeling the median forecast error dependent on the available set of attributes, the point forecast performance is considerably improved as a side effect of prediction interval modeling. This can be considered as dynamic elimination of forecast bias in these models. The results of this study also conform to results obtained by [69]. Yet, the improvement obtained by quantile regression models over clustering based models is considerably greater in the experiments conducted here.

5.6.3. Confidence Level Results

To perform a more comprehensive evaluation on the introduced prediction interval modeling methods, we provide the results for a range of major confidence levels, i.e. 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 0.95. Figure 5.10 depicts the trends of Reliability and Figure 5.11 depicts the trends of Reliability$^{0.95}$ (which uses Coverage$^{0.95}$ rather than

the initially measured Coverage). In addition, a detailed report of observed coverage measures for three confidence levels of 10%, 50% and 95% is provided in Table 5.3. The first column in each section of the table reports the average distance between a missed case and the edge of its forecast interval (δ as defined in subsection 3.2). There is no considerable bias observed in results regarding the occurrence of missed forecasts on the left or right side of the intervals.

Figure 5.12 shows the change of prediction interval width obtained by various models using a range of confidence levels. To perform the final skill score evaluations, one should note that the reliability would have to be measured in terms of the average distance of observations to the prediction interval boundary (i.e. $\bar{\Delta}_{\mathrm{M}}^{\alpha}$). Hence, in Figure 5.13 this measure is projected. Finally, in Figure 5.14 the overall skill of the various methods are compared over the selected range of confidence levels.

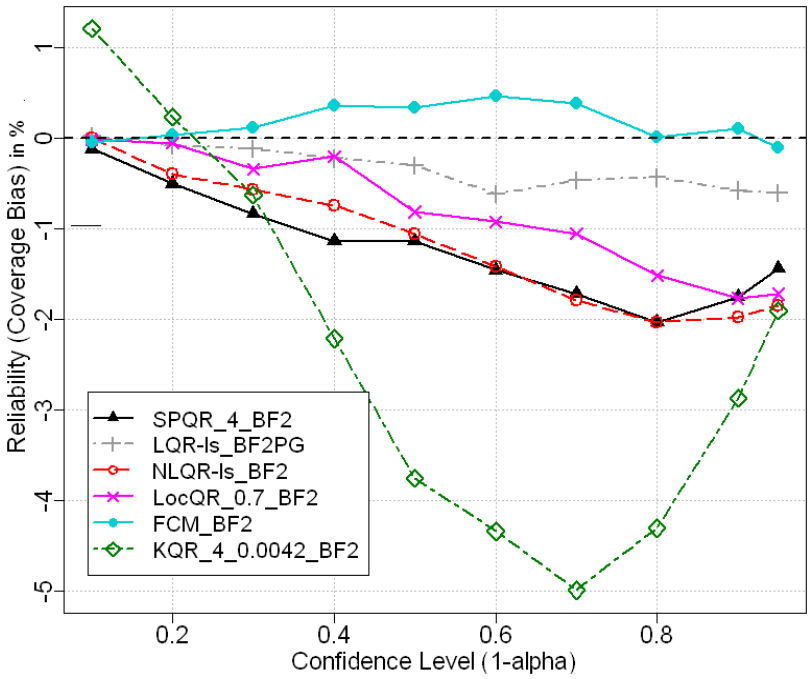


Figure 5.10. Comparison of Reliability between various methods over confidence levels

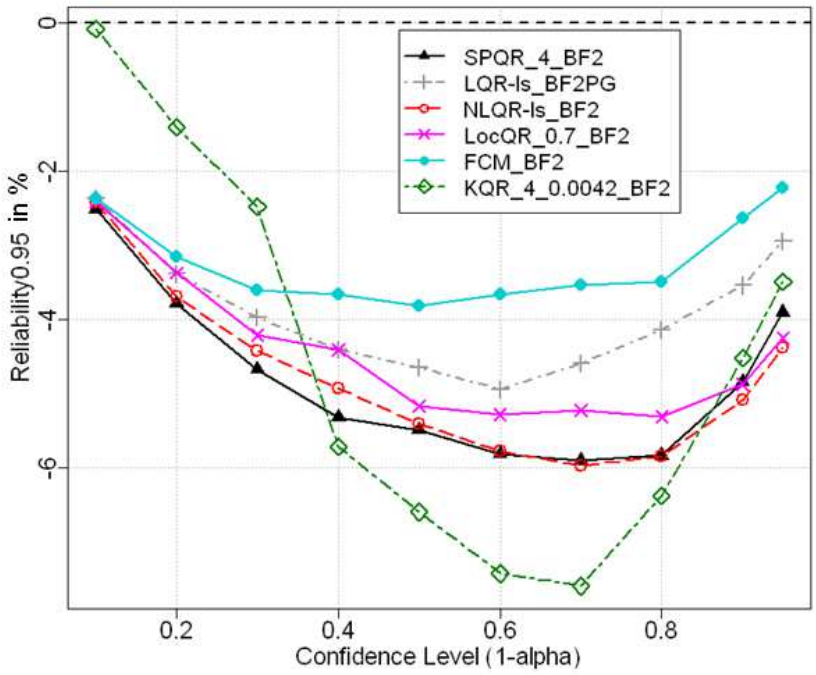


Figure 5.11. Comparison of Reliability0.95 between various methods over confidence levels

Table 5.3. Detailed coverage and miss ratio observations in test for three confidence levels

	$(1-\alpha)$=0.95				$(1-\alpha)$=0.5				$(1-\alpha)$=0.1			
Algorithm	Avg. δ (°C)	Miss (left)%	Hit (center)%	Miss (right)%	Avg. δ (°C)	Miss (left)%	Hit (center)%	Miss (right)%	Avg. δ (°C)	Miss (left)%	Hit (center)%	Miss (right)%
SPQR	0.70	3.3	93.6	3.2	1.06	25.8	48.9	25.3	1.35	45.5	9.9	44.6
LocQR	0.75	3.4	93.5	3.2	1.11	26.8	49.2	24.0	1.41	46.8	10.0	43.2
NLQR	0.78	3.4	93.2	3.4	1.12	25.8	48.9	25.3	1.42	45.3	10.0	53.7
KQR	0.82	3.4	93.1	3.5	1.20	28.4	46.2	25.4	1.55	46.3	11.2	42.5
LQR	0.82	2.8	94.4	2.9	1.22	25.2	49.7	25.1	1.55	45.1	10.0	54.0
FCM	1.08	2.7	94.9	2.4	1.62	24.8	50.3	24.9	2.03	44.9	10.0	45.2
Base-Month	1.11	2.5	95.1	2.4	1.82	24.6	50.7	26.6	2.32	44.8	10.3	44.9

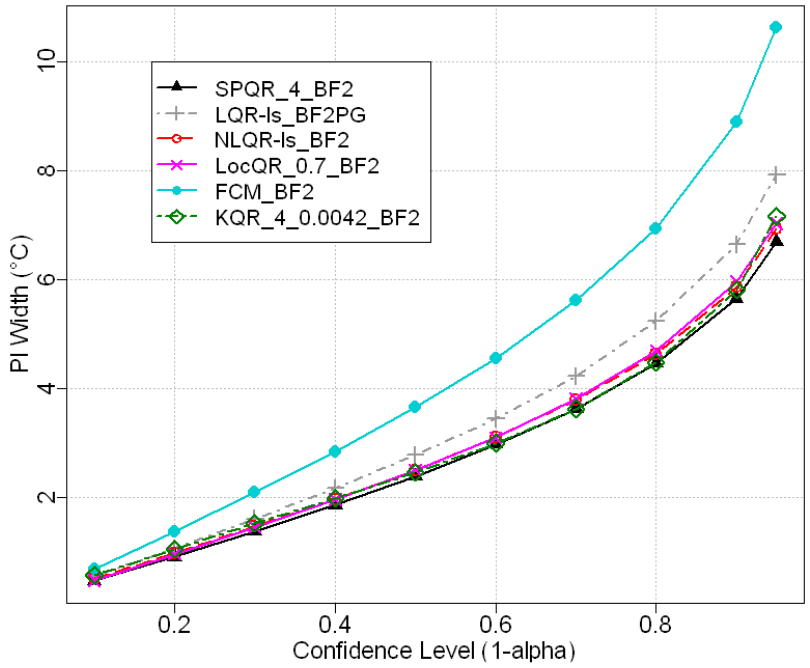


Figure 5.12. Comparison of prediction interval width between various methods over

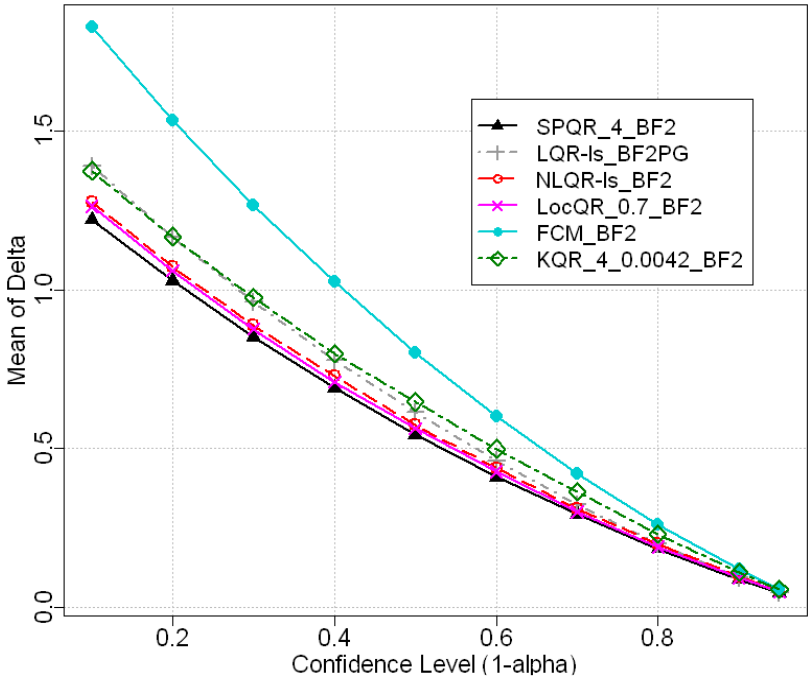


Figure 5.13. Comparison of $\overline{\Delta}_M^{\alpha}$ between various methods over confidence levels

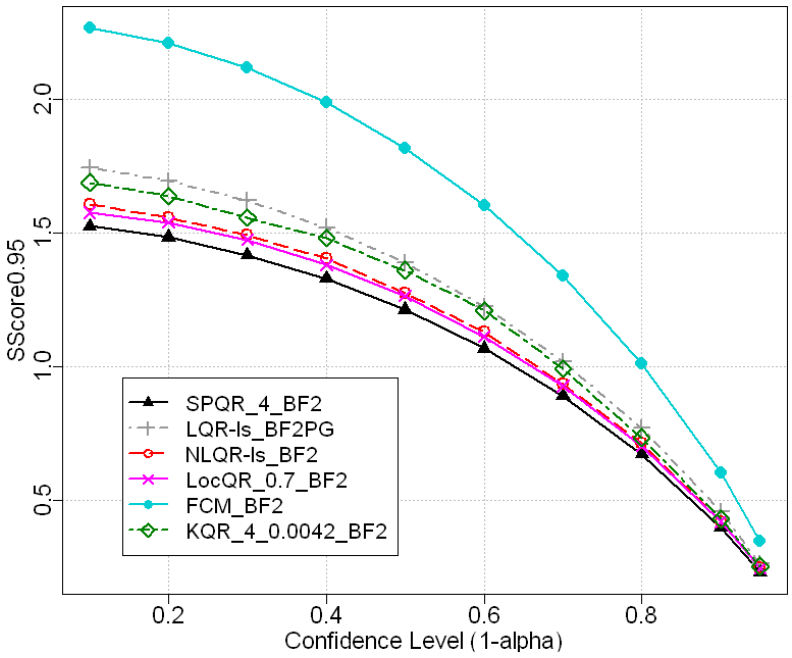


Figure 5.14. Comparison of $SScore_M^{0.95}$ between various methods over confidence levels

The results obtained from the range of confidence levels confirm superior skill of the SPQR method based prediction interval forecasts. This model outperforms other methods both in terms of sharpness and reliability of the provided prediction intervals.

Experiments reveal that the KQR algorithm suffers from lower reliability of prediction intervals although its sharpness is comparable with those of other quantile regression algorithms. However, it is observed in the experiments that the KQR algorithm ranks as the best or second best quantile regression method in terms of overall skill when applied to smaller feature sets including C3 and BF1. This likely demonstrates the higher capability of this algorithm to handle lower dimensionality quantile learning problems.

A possible explanation for the superior performance of quantile regression models over clustering methods is the fact that the forecast error information is directly utilized (in the objective loss function) by the single phase optimization procedure involved in these methods. In contrast, clustering-based methods (including FCM) determine the clusters of forecast cases by an optimization procedure that does not exploit the forecast error, but is based solely on the predicted weather attributes (unsupervised learning). The forecast error information is used only later, in the second phase of distribution fitting.

5.7. Conclusions

Major quantile regression methods including kernel quantile regression and clustering based methods were applied for prediction interval modeling on a data set of

NWP forecasts. These models extend the raw point predictions of the forecasting system into interval forecasts that intrinsically communicate the expected forecast uncertainty to the users. A key analysis for skill score evaluations was taken into consideration in test experiments. The roles of parameters and various available features applied in quantile regression models were investigated in the experiments. The results demonstrated the superior performance of quantile regression models and specifically the spline quantile regression. Prediction interval models obtained from the hybrid method of clustering and KQR can also outperform other models when using low dimensional feature spaces but can only get close to the best model with higher dimensional feature sets. All QR models considerably outperform the clustering-based models in terms of forecast skill. However, it should be noted that clustering models have a higher reliability and can model the entire probabilistic distribution of a forecast in a single model.

Chapter 6
Time Series Approaches to Uncertainty Modeling

This chapter turns the attention of the uncertainty modeling problem into consideration of temporal features of the point forecast errors. In this regards, time series analysis approaches that can provide an estimation of the expected variance of the forecast into the future are investigated. We also consider time series models that focus on the variance of the density forecast rather than only focusing on the mean of this density for future time steps. Prediction intervals obtained from these time series models are then compared with interval forecasts from clustering and quantile regression methods discussed in the previous chapters.

6.1. Introduction

Weather forecast accuracy records are a valuable source of knowledge about the systematic and chaotic behavior of prediction error. By utilizing such information useful models can be obtained that are capable of predicting the uncertainty of system outputs. Different learning methods including clustering and quantile regression are studied as modeling approaches for this purpose in the previous chapters. However, due to the intrinsic temporal quality of the forecasts gained from Numerical Weather Forecasting (NWP) systems, time series modeling can be considered as a potentially suitable approach for uncertainty modeling. Forecast errors of different weather attributes (e.g. temperature and wind speed) are recorded in consecutive time steps (e.g. hourly) in each station which can be considered as a time series. The primary goal of the time series modeling performed in this study is to obtain a temporal model for forecast uncertainty along with the expected value of target. This model can then in turn provide prediction intervals for the target attribute of interest with any desired level of confidence in

different time steps into the future. In this work, different univariate time series modeling methods including ARIMA models and also heteroscedastic models for conditional variance are empirically investigated to obtain prediction intervals for weather forecasts. The quality and skill of time series uncertainty forecasts for 1-hour-ahead up to 10-days-ahead are evaluated and also compared against clustering and quantile regression forecasts using the prediction interval evaluation framework.

Time series models have been broadly studied and applied in various weather forecast problems such as climatological forecasting [85][55] and short to medium-range forecasting [24][38][52]. As another motivation for applying statistical methods in our problem, Wilks [86] notes that these methods "…are still viable and useful at very short lead times (hours in advance), or very long lead times (weeks or more in advance), for which NWP information is not available with either sufficient promptness or accuracy, respectively.". Hence, we are motivated here to apply appropriate time series models for the purpose of uncertainty modeling and compare the accuracy of the forecasts obtained against the other uncertainty prediction models investigated in this study.

In [10] a non-structural time series modeling approach is taken to forecast daily average temperatures for weather derivative applications. Due to the crucial significance of forecast uncertainty in the weather derivative market, GARCH models are also used to provide estimations of target densities into the future. The accuracy of point forecasts obtained from this time series model is compared against benchmark methods and also an NWP model. Results of this study confirm better performance of the autoregressive time series model when compared to the benchmark methods and also show that this model can even outperform the NWP model in longer horizons (i.e. leads bigger than 8 days). The capacities of these models in terms of volatility prediction are not however broadly evaluated and are not compared versus other available uncertainty modeling methods.

Franses *et al.* [24] use GARCH models to capture volatility clustering and obtain a univariate model for weekly mean temperatures. Time series models for point and density forecasts of daily temperature are compared with ensemble predictions from an NWP atmospheric model in [78] and [79]. An AR-GARCH model is fitted to the series and utilized to provide forecasts of mean and variance for given horizons based on a Gaussian distribution assumption for residuals. The ensemble mean forecast outperforms the other methods in terms of accuracy both for point and quantile forecasting. Yet, the time series model can provide forecasts that are even more accurate than the NWP forecast for

horizons longer than 6 days. However, in this study the ensemble model-based uncertainty forecasts are assumed to be inaccessible and uncertainty models are obtained using solely the historical records of the point forecasting system. In this context, the study performed here is novel as it compares major time series models along with other uncertainty modeling methods of clustering and quantile regression by focusing on the historical performance of an empirical weather forecasting application.

Accordingly, the time series models learned from the available system performance history are to provide an estimate of forecast uncertainty (along with the expected value) which can then be reported in the format of prediction intervals. For this purpose, two main groups of time series models are employed. The first set of methods applies the theoretical statistical analysis of residual variance of seasonal ARIMA models to obtain variance forecasts along with expected mean predictions. In the second set, explicit autoregressive models of variance are employed to train time series models of conditional variance over the forecasts. These models are then subject to analytical and comparative study along with other models of forecast uncertainty, namely clustering-based and quantile regression-based models.

6.2. Time Series Modeling Essentials

Similar to other machine learning approaches, time series modeling involves learning a model using available training data and later evaluating the quality of the model by test data. However, there is a wide range of models to choose from when fitting a time series model in the training phase. Each of these models would be appropriate for different types of time series depending on the process which the data is generated from. Hence, the general procedure of time series modeling involves the three phases of a) model specification, b) model fitting and c) model diagnostics [19].

As the first step “model specification” is required that investigates various characteristics of the time series to determine the most appropriate model to be fitted to data. Next, the specified model is fitted to the available time series data by estimating the parameters defined in the model. Classical techniques such as Least Squares Estimation (LSE) or Maximum Likelihood Estimation (MLE) are used for this purpose. Finally, in the third step the learned model undergoes diagnostic analysis to determine whether there has been a shortcoming in either of the two previous steps. Often, the last step would advise that the initially specified model should be redefined and/or fitted again for couple of cycles until a proper time series model is lastly obtained. In this study, we aim to

obtain models of "temperature forecast error time series". After providing basic background and definitions in this section, major analysis and results from each of the above mentioned modeling steps are explained in the following sections.

6.3. Definitions and Processes

A time series is essentially a stochastic process consisting the sequence of random variables $\{Y_t : t = 0, \pm1, \pm2, \pm3, \dots\}$. For the Y_t series there are some characteristics that describe major aspects of the series. These include the mean, variance and autocorrelation of the process:

$$\mu_t = E(Y_t) \quad for\ t = 0, \pm1, \pm2, \dots \quad (6.1)$$

$$\gamma_{t,s} = Cov(Y_t, Y_s) \quad (6.2)$$

$$\rho_{t,s} = Corr(Y_t, Y_s) = {\gamma_{t,s}}\Big/{\sqrt{\gamma_{t,t}\gamma_{s,s}}} \quad (6.3)$$

where $s = 0, \pm1, \pm2, \pm3, \dots$ and $\gamma_{t,s}$ is the autocovariance function. Function $\rho_{t,s}$ is the autocorrelation function which provides a unit-less measure of (linear) dependence between the two random variables. Hence, *t* and *s* are arbitrary time indexes that can be chosen to have any value in the above equations.

Random Walk

A "Random Walk" as the simplest form of a time series is represented as follows [25]:

$$Y_t = Y_{t-1} + e_t \quad (6.4)$$

where e_t is a white noise stochastic process with zero mean and σ_e^2 variance and the random variables $e_1, e_2, \dots$ are independent and identically distributed (i.i.d.). This defines the time series in a way that each new sample in the series is the result of a random change from the previous value. To investigate the above metrics for this series we have:

$$\mu_t = E(e_1 + e_2 + \cdots + e_t) = 0 \quad for\ all\ t \quad (6.5)$$

$$Var(Y_t) = \gamma_0 = Var(e_1 + e_2 + \cdots + e_t) = t\sigma_e^2 \quad (6.6)$$

which represent the mean and variance of the Y_t random variable.

Stationarity

A critical concept in the study of time series is "stationarity". The idea behind this concept is that "the probability laws that govern the behavior of the processes do not change over time. In a sense the process is in statistical equilibrium" [19]. Process $\{Y_t\}$ is

"strictly stationary" if the joint distribution of $Y_{t_1}, Y_{t_2}, \dots, Y_{t_n}$ is the same as the joint distribution of $Y_{t_1-k}, Y_{t_2-k}, \dots, Y_{t_n-k}$ for all choices of *t* and *k*. One common way of examining the stationarity of a series is to check whether its covariance function only depends on the time lag (i.e. *k*=*s*-*t*) and not the actual time positions (i.e. *t* or *s*). For example, the random walk series is not stationary since its covariance function:

$$\gamma_{t,s} = \sum_{i=1}^{s} \sum_{j=1}^{t} Cov(e_i, e_j) = t\,\sigma_e^2 \quad (6.7)$$

$$\rho_{t,s} = \sqrt{\frac{t}{s}} \quad (6.8)$$

suggests that the covariance (and hence the correlation) function is dependent on the actual time position meaning that the correlation characteristics of the series actually changes during time in contrary to the definition of stationarity.

Trend in Time Series

In a stationary process the mean function must be a constant function of time. Due to the regular inclination of the Northern Hemisphere toward the sun a seasonal trend is naturally expected for the forecasted temperature values. However, here the focus of the analysis is on the temperature *error* series which does not necessarily follow the same cyclical trend. By considering a seasonal trend function a possible model can be:

$$Y_t = \mu_t + X_t \quad (6.9)$$

where μ_t is a deterministic annual periodic function i.e. $\mu_t = \mu_{t-12}$. The general assumption for monthly seasonal data considers 12 constants for the expected mean value for each month:

$$\mu_t = \begin{cases} \beta_1 & for\ t = 1,13,25,\dots \\ \beta_2 & for\ t = 2,14,26,\dots \\ \vdots & \\ \beta_{12} & for\ t = 12,24,36,\dots \end{cases} \quad (6.10)$$

Figure 6.1 shows the box plots of such seasonality trends as reported in Table 6.1, sometimes called a seasonal mean model, for forecasted temperature and error series. Error series also show some considerable level of seasonality in the trend although they are less substantial when compared to that of the forecasted temperature trend.

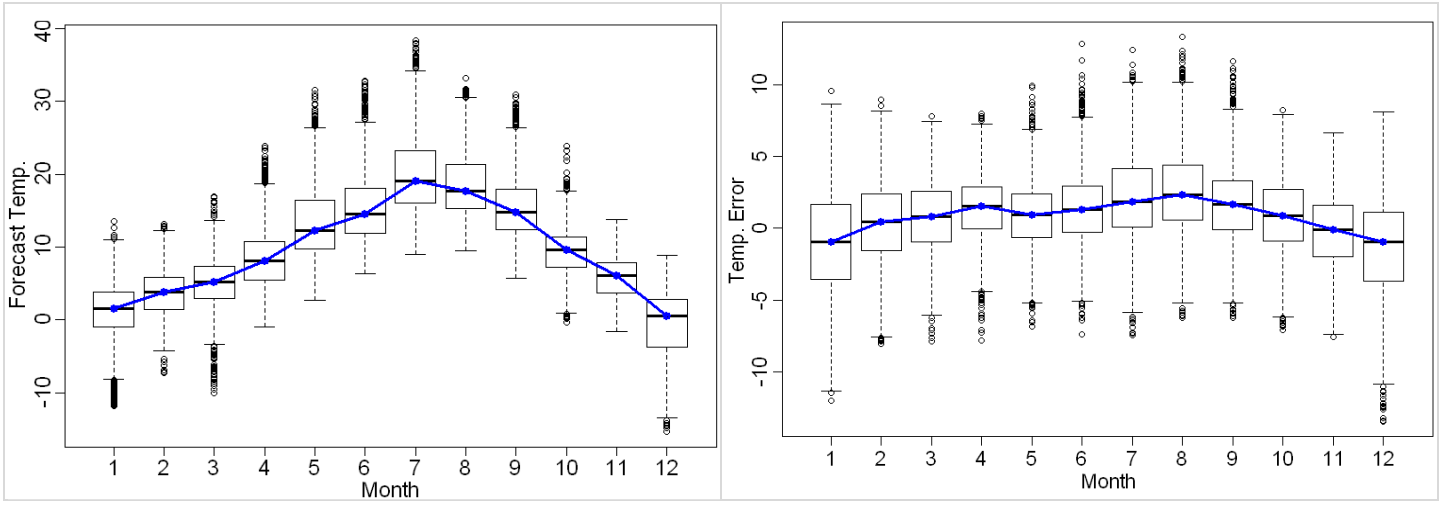


Figure 6.1 Seasonality trend of the (left) forecast temperature series and (right) the forecast error time series

Table 6.1. Seasonal mean model parameters for forecast temperature and error

Month	$\widehat{\boldsymbol{\mu}}_{\boldsymbol{t}}$ (temp.)	$\widehat{\boldsymbol{\mu}}_{\boldsymbol{t}}$ (error)
January	1.6	-0.97
February	3.8	0.47
March	5.2	0.83
April	8.1	1.52
May	12.3	0.91
June	14.5	1.29
July	19.1	1.86
August	17.7	2.35
September	14.8	1.67
October	9.6	0.90
November	6.1	-0.13
December	0.6	-0.98

For a detailed analysis on the accuracy of such constant mean estimation of the trend one can refer to subsection 3.2 of [19]. To achieve a better smooth transition between the time periods the seasonal trend can be modeled with sinusoidal curves:

$$\mu_t = \beta_0 + \beta_1 \, cos(2\pi f t) + \beta_2 \, sin(2\pi f t) \quad (6.11)$$

where the β parameters are estimated by regression. Figure 6.2 shows a fitted cosine trend (parameters shown) to the temperature forecast error along with the observed errors as points. One can notice that the general fluctuations in error are captured by the trend curve. Details on the reliability of the regression estimates can be found in [25].

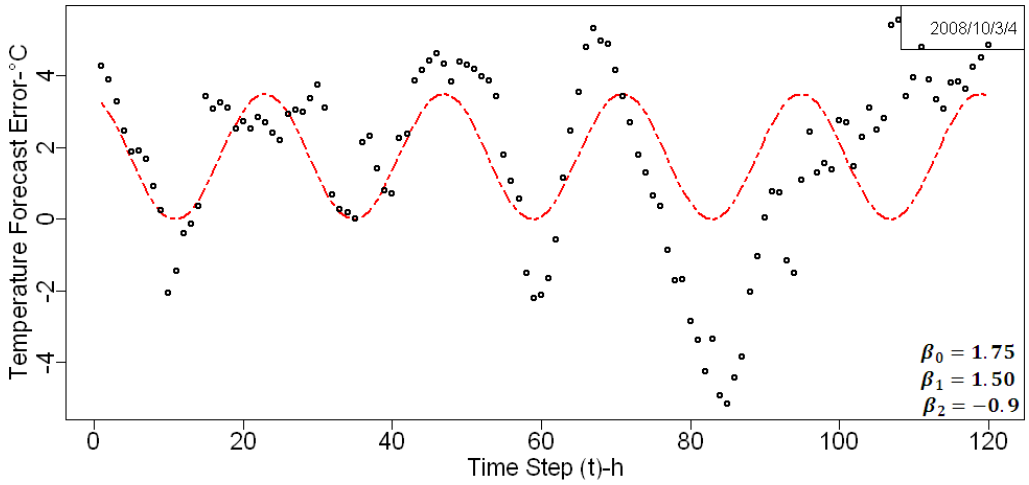


Figure 6.2. Cosine trend fitted to 5 days of temperature error

When examining the residuals of a trend model we have:

$$\hat{X}_t = Y_t - \hat{\mu}_t \quad (6.12)$$

where $\hat{X}_t$ is the residual of the *t*th observation. Using a least squares fit will automatically result in a zero mean residual and it can be standardized by diving over the standard error. Residual analysis of both models described above reveals that there is dependence and non-normality evident in the residuals of the respective models which signals the need for further complex model fitting strategies discussed subsequently.

General Linear Process

Using this model $\{Y_t\}$ is represented using a weighted linear combination of present and past white noise terms [74]:

$$Y_t = e_t + \Psi_1 e_{t-1} + \Psi_2 e_{t-2} + \cdots, \qquad \sum_{i=1}^{\infty} \Psi_i^2 < \infty \quad (6.13)$$

A typical example is the case where the weights are an exponentially decaying sequence:

$$\Psi_j = \phi^j, \quad -1 < \phi < 1 \quad (6.14)$$

for this case we can easily calculate the characteristic measures as:

$$E(Y_t) = 0, \quad Var(Y_t) = \frac{\sigma_e^2}{1-\phi^2}, \quad Corr(Y_t, Y_{t-k}) = \phi^k \quad (6.15)$$

Hence, this process is stationary and one can obtain a nonzero mean process by adding the μ (trend element) to Equation (6.13).

Moving Average Process

Using a finite number of non-zero weights in the general linear process the moving average process is obtained [19]:

$$Y_t = e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (6.16)$$

where q is the order of the process MA(q). It can be shown by calculations that for this process we have:

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2 + \cdots + +\theta_q^2)\sigma_e^2 \quad (6.17)$$

$$\rho_k = \begin{cases} \frac{-\theta_k + \theta_1\theta_{k+1} + \theta_2\theta_{k+2} + \cdots + \theta_{q-k}\theta_q}{1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2} & for\ k = 1,2,\ldots,q \\ 0 \quad for\ k > q \end{cases} \quad (6.18)$$

hence, there is a cut off after lag q.

Autoregressive Process

When the random process is the regression of its previous values we obtain the definition of *p*th order autoregressive process as follows [15]:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \quad (6.19)$$

For example, when considering AR(2), the second-order autoregressive, we have:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \quad (6.20)$$

where e_t is assumed to be independent of Y_{t-1} and Y_{t-2}. For investigation of stationarity we consider the *AR characteristic polynomial* and equation:

$$\phi(x) = 1 - \phi_1 x - \phi_2 x^2 = 0 \quad (6.21)$$

It can be shown that the process will be stationary if and only if the absolute values of roots for the characteristic equation exceed 1 or in other words the roots should lie outside the unit circle in the complex plane. To gain the autocorrelation function, we first obtain the autocovariance function after multiplying Equation (6.19) by Y_{t-2} and taking expectations:

$$\gamma_k = \phi_1 \gamma_{k-1} + \phi_2 \gamma_{k-2} \quad for\ k = 1,2,3,\ldots \quad (6.22)$$

which are usually called the *Yule-Walker equations*. Various values of ρ_k can be obtained by dividing this equation by γ_0, setting k and ρ_0 equal to 1 and calculating the autocorrelation values for higher lags successively. Details of these calculations and also methods to directly calculate the autocorrelation values are provided in [25]. It is also shown that with complex root for the Yule-Walker equations the correlogram exhibits a sine wave shape with a damping factor which is dependent on the roots too.

Also note that an autoregressive model can be expressed as a general linear process as defined in Equation (6.13). By using the recursive definition of an AR model we can get the values for Y_{t-1}, Y_{t-2}, etc. and by substituting these into the original equation one will obtain a general linear process version of the original process.

6.3.1. Mixed Autoregressive Moving Average Process

A more general form for a time series can be obtained by assuming a series being partly autoregressive and partly moving average [19]:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} \quad (6.23)$$

where the process is called ARMA(p,q). It should be noted that we may get the same autocorrelation functions using different values of θ. Hence, given an autocorrelation function there may not be a unique MA model for it. To address this issue of invertibility we first introduce the re-expression of an MA process as an infinite-order AR process by substituting for consecutive values of e_{t-1} and defining the *MA characteristic polynomial* [19]. It can be shown that the MA(q) model is invertible if and only if the roots of the MA characteristic equation exceed 1 in modulus. Both stationarity and invertibility are required for the ARMA model.

6.3.2. Nonstationary and ARIMA Process

When there are not sufficient reasons to assign a deterministic trend for a series (e.g. just a linear increase in a segment of the series), one would have to use nonstationary models to fit the data which consider stochastic trends. It can be shown that using different sets of assumptions the first or second difference of many non-stationary models, leads to a stationary process [19].

When the dth difference $W_t = \nabla^d Y_t$ is a stationary ARMA(p,q) process, Y_t is identified as an integrated autoregressive moving average process i.e. ARIMA(p,d,q) where d is considered 1 or 2 practically. For d=1, $W_t = Y_t - Y_{t-1}$ and for higher values of d this transformation is repeated d-1 times over W_t. When there are no autoregressive or moving average terms in the process it is denoted as ARI and IMA, respectively. By substituting associated difference values in the ARIMA formulation the equivalent ARMA model can be obtained which is of course non-stationary (a unit root exists for the characteristic polynomial).

To obtain stationarity in a series which has increased dispersion for higher values, the logarithm transformation can be used. The power (Box-Cox) transformations are another

alternative to obtain normality and stationarity [86]. It can also be shown that when Y_t is a relatively stable percentage change from Y_{t-1}, a log transformation followed by a first difference can provide a stationary process known as the *returns* series in the financial domain.

6.4. Time Series Models for Temperature Forecast Error

6.4.1. Model Specification

For basic analysis of the time series, 5 days of hourly recorded temperature (at 2m) forecast errors starting from first day of June, 2007 by the WRF NWP system are plotted in Figure 6.3 for the a weather station in Hope, BC, Canada. Also for a broader look at the characteristics of this time series a 45 day long slot of the same series is depicted in Figure 6.4. In these figures, high autocorrelation and seasonality qualities of the series are evident.

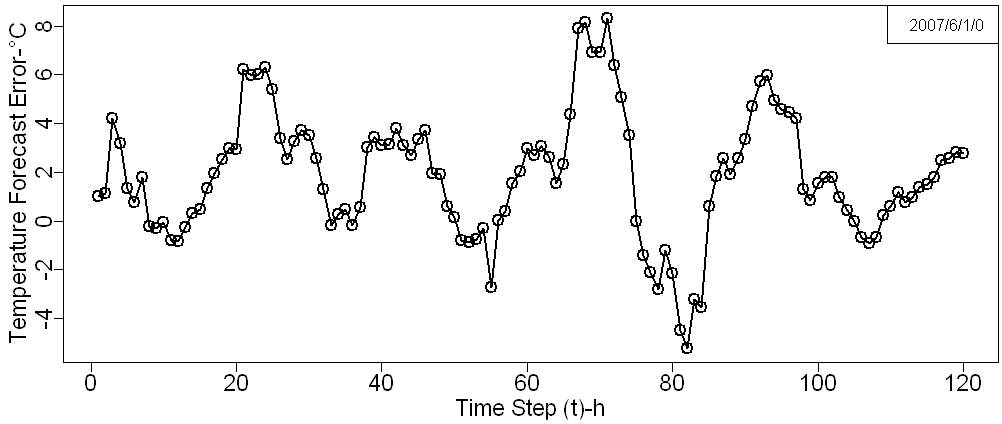


Figure 6.3. Temperature forecast error time series

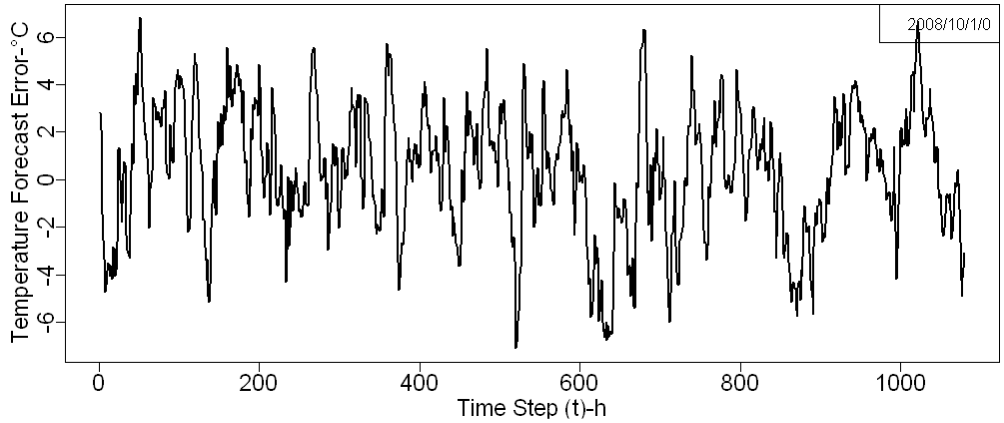


Figure 6.4. 45 days long series of temperature forecast error

To have a detailed look at the dependency of the current temperature error and forecast errors in the past hours, scatter plots of different lags are represented in Figure

6.5. Strong autocorrelation is apparent between the current temperature error and the system error observed in the previous hour. Correlation with a lesser extent is also observable between the current forecast error and that of the same hour on the previous day. However, the 12-hour lag shows a rather different behavior as there is less correlation when compared to the 24-hour lag.

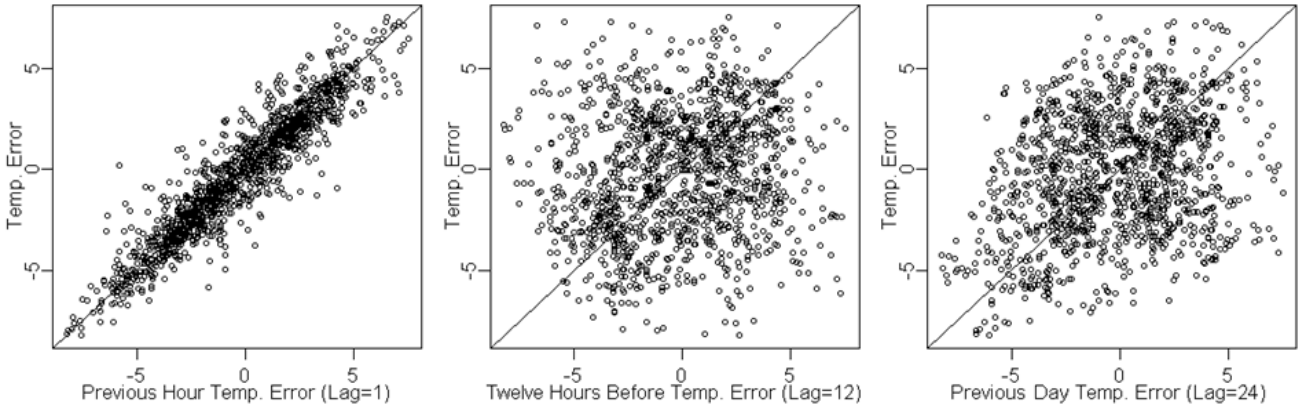


Figure 6.5. Autocorrelation of the forecast error series in different lags

A closer look at the autocorrelation measures in various lags of 1, 12 and 24 hours is provided in Figure 6.6. These values are measured independently for each month rather than calculating them using the whole time series. Unlike the 1 and 24 hour lags, 12 hour lag values exhibit a considerably higher variance in different months and maintain this pattern through the different years.

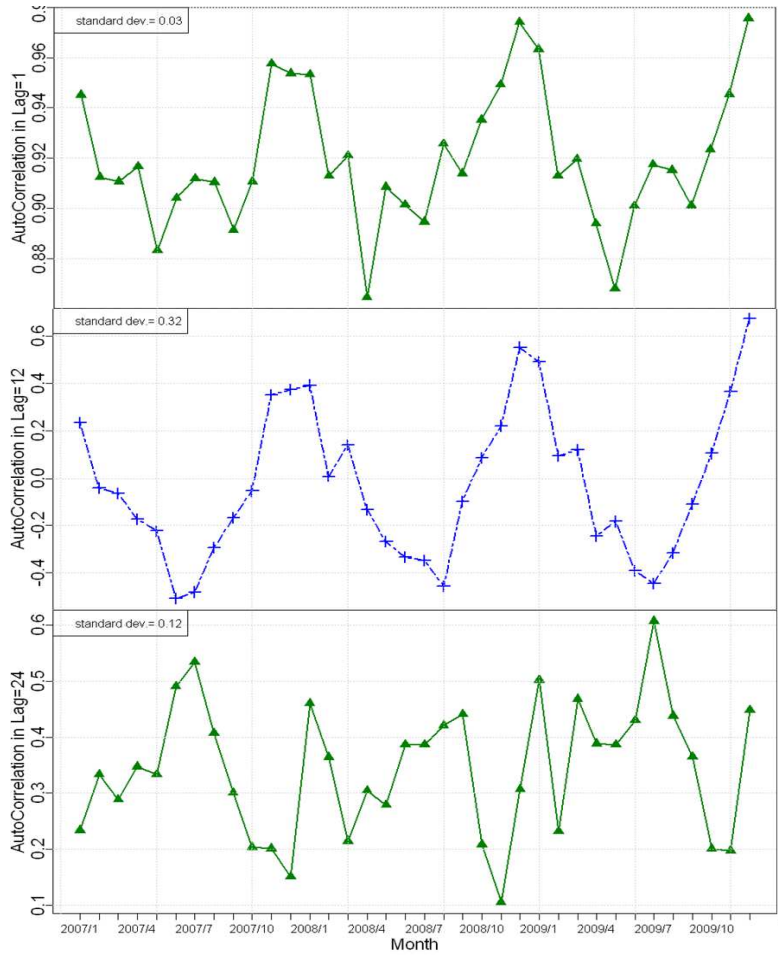


Figure 6.6. Autocorrelation for three different lags (i.e. 1, 12 and 24 hours) projected over month

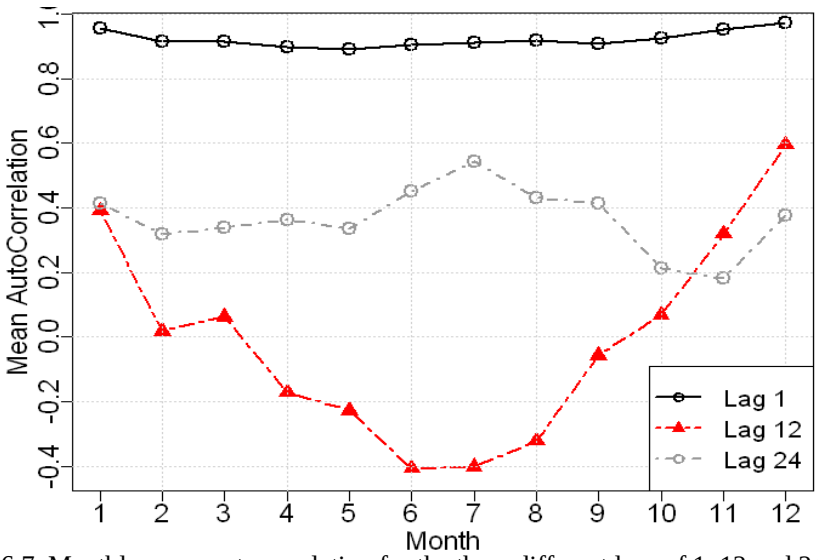


Figure 6.7. Monthly mean autocorrelation for the three different lags of 1, 12 and 24 hours

Figure 6.7 demonstrates this condition with lag 12 values fluctuating between -0.4 in June and 0.6 in December. Such information may suggest special considerations such as

considering two independent models for cold and warm seasons in the model specification phase.

6.4.2. Sample Autocorrelation Function

The available recorded time series data can be used to calculate the sample autocorrelation values for different lags:

$$r_k = \frac{\sum_{t=k+1}^{n}(Y_t-\bar{Y})(Y_{t-k}-\bar{Y})}{\sum_{t=1}^{n}(Y_t-\bar{Y})^2} \quad (6.24)$$

The r_k values can then be checked to see whether they follow a characteristic pattern for a common ARMA model with specific parameters. For instance, for MA(q) processes the autocorrelation function (ACF) is zero for lags beyond q. Figure 6.8 shows the sample correlation function for the temperature forecast error time series in the Hope station for up to 72 hour lags. In this correlogram there is a high correlation observed in the first few lags which gradually declines reaching its minimum in the 12th hour lag and increases again in the next 12 hours. This pattern is maintained through the next days with a damped property.

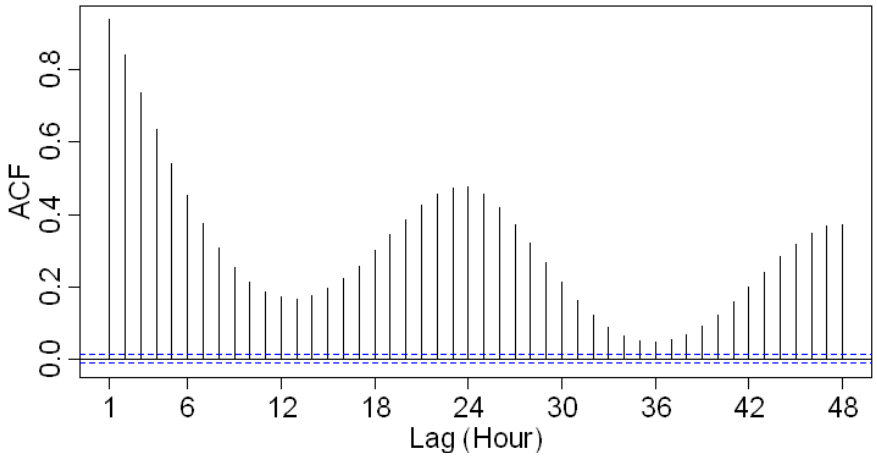


Figure 6.8. Sample correlogram for up to 2 days back

The sample correlations are subject to sampling variation yet its properties are not easily obtained as it is a ratio of quadratic functions with dependent variables. Results from simulations provide methods for computation of the sampling distribution. For large n, r_k is assumed to be approximately normally distributed with mean ρ_k and a variance that is calculated using a formula defined in [19] which is inversely proportional to the number of samples. Consequently, the variance of the sampled autocorrelation values can be obtained by these formulas and then considered for hypothesis testing purposes.

6.4.3. Sample Partial and Extended Autocorrelation Functions

For an AR(p) model the autocorrelation function does not become zero after an specific lag. Hence a different function is required to decide about the proper order for AR models. Therefore the partial autocorrelation function (PACF) is defined which considers the correlation between Y_t and Y_{t-k} after removing the effect of the intervening variables $Y_{t-1}, \dots, Y_{t-k+1}$. Assuming $\{Y_t\}$ as a normally distributed series:

$$\phi_{kk} = Corr(Y_t, Y_{t-k} | Y_{t-1}, Y_{t-2}, \dots, Y_{t-k+1}) \quad (6.25)$$

where ϕ_{kk} is the partial correlation at lag k. It can be shown that for an AR(p) process the partial autocorrelation function cuts off after lag p. For an MA(q) process the partial autocorrelation function will also decay exponentially to zero.

To estimate the function based on an observed time series the following recursive equation can be used [19]:

$$\phi_{kk} = \frac{\rho_k - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_{k-j}}{1 - \sum_{j=1}^{k-1} \phi_{k-1,j} \rho_j} \quad (6.26)$$

where $\phi_{k,j} = \phi_{k-1,j} - \phi_{kk} \phi_{k-1,k-j}$ for j=1,2,…,k-1 and $\phi_{1,1} = \rho_1$. In practice, the ρ values are replaced r values. It is also shown that for an AR(p) process the sample partial autocorrelations at lags greater than p are approximately normally distributed with zero mean and variance $1/n$.

It is difficult to identify mixed ARMA models using sample ACF and PACF. The extended autocorrelation function (EACF) is known to be a good tool for this purpose with large sample sizes [13]. This method uses a finite sequence of regressions to filter out the AR part of a mixes ARMA model to obtain a pure MA process that enjoys the cutoff property in its ACF. The EACF information is summarized by a table which reports the sample correlations of the autoregressive residuals (assuming different AR and MA orders) which are significantly different from zero [82]. In this table an ARMA(p,q) process will theoretically have a triangle of zeros with the left vertex matching the proper orders of the model.

Figure 6.9 shows the sample partial autocorrelation function for the temperature error time series. This graph is a strong indication of an autoregressive model with order 2 since the first two lag partial autocorrelations are significantly different from zero (the dashed horizontal lines represent the critical values for significance test).

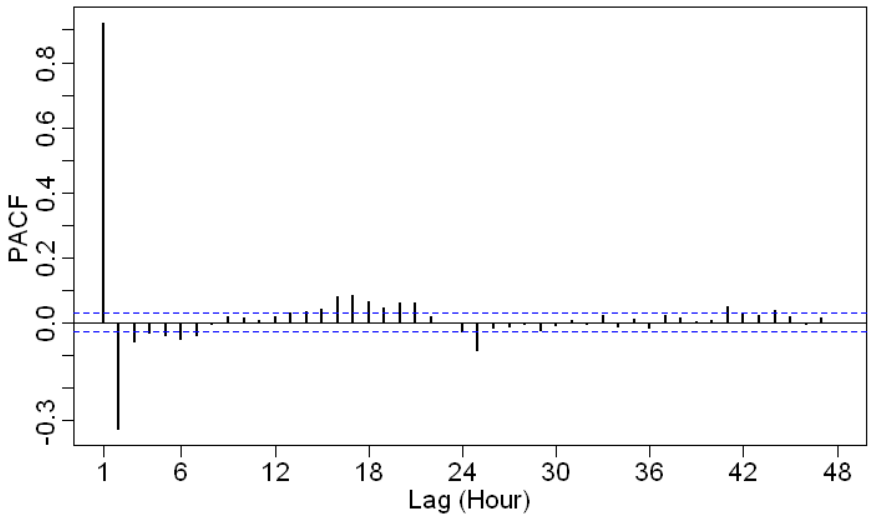


Figure 6.9. Sample partial autocorrelation function for the past two days

The summarized analysis of the extended autocorrelation function is provided in Table 6.2. Entries of "x" represent autoregressive residuals whose sample autocorrelation is significantly different from zero. Please refer to [19] for detailed definition of EACF. The upper left hand zero element which is highlighted clearly suggests that an ARMA(2,2) model would be appropriate for the series.

Table 6.2. Sample EACF for the temperature error series

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	x	x	x	x	x	x	x	x	x	x	x	x	x
2	x	x	**o**	x	x	x	x	x	o	o	o	x	o	x
3	x	x	x	x	x	o	x	o	o	o	o	o	o	x
4	x	x	o	x	o	o	o	o	o	o	o	o	o	o
5	x	x	x	x	o	o	x	o	o	o	o	o	o	o
6	x	x	x	x	x	o	o	o	o	o	o	o	o	o
7	x	x	x	x	x	o	o	o	o	o	o	o	o	o

In the specification process we should be cautious that many series are nonstationary. The ACF of such series typically shows large values of autocorrelation that fail to die out as early as expected. Considering the ACF of the forecast error series in Figure 6.8 all values are significantly far from zero.

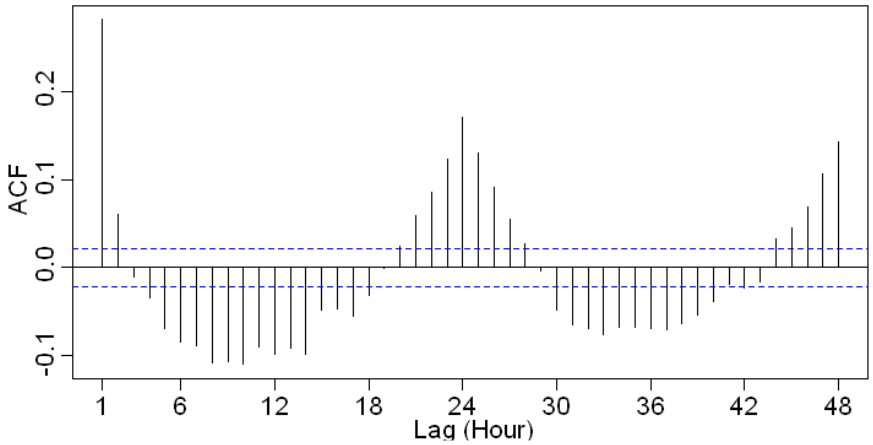


Figure 6.10. Correlogram for first difference of the monthly seasonality removed series

The sample ACF of the first differences of this series is provided in Figure 6.10. It must be noted that over-differencing can result in a nonstationary or noninvertible series and hence must be avoided by differencing in succession and applying parsimony (models should be simple, but not too simple).

6.4.4. The Dickey-Fuller Unit-Root Test

To statistically analyze the nonstationarity of a given series hypothesis testing can be used. Under the null hypothesis that $\{Y_t\}$ is difference nonstationary it can be shown that the AR characteristic polynomial of a properly modified equation of the series, which is an AR(k) process, will have a unit root [19]. The Augmented Dickey-Fuller (ADF) test statistic is the t-statistic of the estimation of a coefficient named a using least squares regression where a=0 corresponds to the null hypothesis of difference nonstationarity [14].

The same test can be used to examine the null hypothesis of a process being linear-trend nonstationary. This can be performed by adding an intercept term and the covariate time in the test's regression model. The ADF test statistic for the seasonality removed temperature error series is estimated -21.7 with the p-value being 0.01 which is an indication (although not very strongly) of the series being stationary.

Another method for selection of orders for an ARMA model is based on minimization of Schwarz Bayesian Information Criterion (BIC) [25]:

$$BIC = -2\,log(maximum\ likelihood) + klog(n) \quad (6.27)$$

where $k = p + q + 1$ for a model with constant term which is included as a penalty function to prefer simple over too complex models. The process minimizing the BIC involves first fitting a high-order AR process with the order determined by minimizing

Akaike's Information Criterion (AIC) and then using residuals as proxies for unobservable error terms and estimating BIC for *k* lags of observations along with *j* lags of the residuals from the high order autoregressive model.

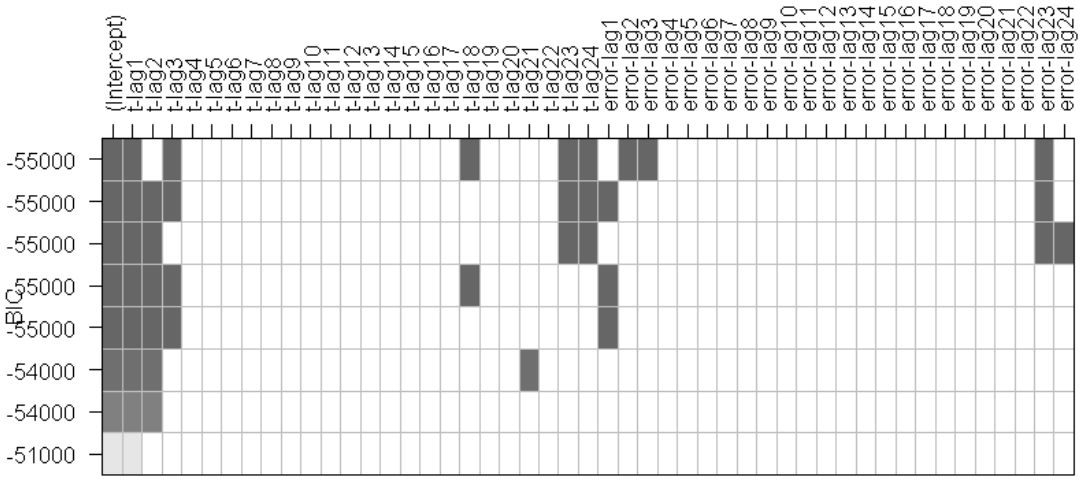


Figure 6.11. Best subset ARMA selection table based on BIC

In Figure 6.11 the results from a BIC-based order selection process are provided. Each row represents a subset ARMA model with orders determined by the shaded cells. These models are ordered according to their BIC. The Best model includes lags 1, 3, 18, 23 and 24 of the time series and lags 2, 3 and 23 of the error process. However, the BIC values are very close for the top models and it is recommended to investigate orders of the other models in further analysis e.g. lag 2 of the time series and lag 24 of the error process.

6.5. ARIMA Model Fitting

6.5.1. The Method of Moments

Assuming that p and q orders for an ARMA model have been already specified, the parameters involved in the model have to be estimated in the next step. In the Method of Moments (MM) theoretical moments of the model are equated with the sample moments and solved to obtain the unknown parameters. For instance, in the general AR(p) case ρ_k values are replaced by r_k estimates in the Yule-Walker equations to obtain estimates of the $\phi_{1..p}$ parameters. For MA and ARMA models the method is more complicated involving quadratic equations that have multiple solutions and only one is invertible and acceptable. It can be shown that this method is only efficient for AR processes and fails for models that have MA processes [19].

6.5.2. Least Square Estimation

The parameters of a model can be obtained using Least Square Estimation (LSE) by minimizing the sum of squares of model error. For example in an AR(1) with constant mean S_c [19]:

$$S_c(\phi, \mu) = \sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 \quad (6.28)$$

is called the conditional sum-of-squares function and is minimized through setting its gradients relative to the parameters equal to zero. It can be shown that the conditional least squares estimation of the general AR model amounts to solving the Yule-Walker equations.

For MA models the objective function will be nonlinear in the θ parameters and hence derivative-based methods are to be replaced by numerical optimization methods such as multivariate Gauss-Newton. The same approach is used for ARMA(*p*,*q*) models by minimizing $S_c(\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q)$.

6.5.3. Maximum Likelihood Estimation

The likelihood function *L* is defined as the joint probability density for actually observing the series in hand. In ARIMA models such function has parameters of $\phi_{1..p}$, $\theta_{1..q}$, μ and σ_e^2. By maximizing the likelihood function with respect to these parameters the observed data are the most likely outcome of the process. The white noise terms are assumed independent and normally distributed with zero mean and each error term can be substituted by its respective model based terms in this joint pdf to obtain the likelihood function. For example in AR(1):

$$f(y_2, y_3, \dots, y_n | y_1) = (2\pi\sigma_e^2)^{-(n-1)/2} exp\left\{-\frac{1}{2\sigma_e^2}\sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2\right\} \quad (6.29)$$

where y_1 is the observed value of the variable Y_1 and so forth. By multiplying this function to the marginal pdf of Y_1 the unconditional sum-of-squares function is obtained:

$$S(\phi, \mu) = \sum_{t=2}^{n}[(Y_t - \mu) - \phi(Y_{t-1} - \mu)]^2 + (1 - \phi^2)(Y_1 - \mu) \quad (6.30)$$

and then applying the logarithm over *S* will result in the log-likelihood function which can be numerically minimized to obtain parameters fitted to the observed series. Since $S(\phi, \mu) \approx S_c(\phi, \mu)$ the parameter estimation should be very similar for large sample sizes. For details on derivation of Maximum Likelihood Estimation (MLE) functions for general ARMA models we refer the reader to [74].

By using the maximum likelihood theory we can obtain the sample variance properties of the estimated parameters. For instance in the AR(1) model we have:

$$Var(\hat{\phi}) = \frac{1-\phi^2}{n} \quad (6.31)$$

Hence, the variance of the ϕ estimator decreases as ϕ approaches ±1. The analysis can also provide correlation between multiple estimated parameters. It can also be shown that the variance in the method of moments is always larger than that of the maximum likelihood estimation. For example for an MA(1) model with θ equal to 0.9, the sd. of its estimation using MM will be more than 5 times larger than that of the MLE method.

Based on the analysis of the temperature forecast error series performed in the previous subsection regarding the results of extended autocorrelation function and BIC analysis, we start fitting ARMA models with orders of 2 and 3. In Table 6.3 the results of parameter estimation using least squares and maximum likelihood estimation are reported for the Hope station. The model has been trained using the first two years of data (i.e. 2007 and 2008) which has about 17000 observations.

Table 6.3. LSE and MLE estimates of the favourable ARMA models for the temperature error series

	ARMA(2,2)		ARMA(2,3)		ARMA(3,2)	
Parameters	Conditional LSE	MLE	Conditional LSE	MLE	Conditional LSE	MLE
ϕ_1	1.63±0.04	0.21±0.45	1.64±0.04	1.63±0.05	1.72±0.15	1.44±0.46
ϕ_2	-0.69±0.04	0.59±0.41	-0.69±0.04	-0.69±0.04	-0.83±0.23	-0.41±0.66
ϕ_3	-	-	-	-	0.06±0.09	-0.11±0.24
θ_1	0.41±0.04	-0.99±0.43	0.42±0.04	0.41±0.05	0.50±0.15	0.23±0.46
θ_2	0.11±0.02	-0.25±0.79	0.11±0.02	0.11±0.02	0.08±0.06	0.17±0.10
θ_3	-	-	0.01±0.02	0.01±0.02	-	-
intercept	-0.07±0.13	-0.07±0.18	-0.07±0.13	-0.07±0.13	-0.07±0.13	-0.07±0.14
σ_e^2	0.99	1.01	0.99	0.99	0.99	0.99

By starting from a more specific model of ARMA(2,3) the estimated parameters for the autoregressive and moving average components tend to be significantly different from zero (meaning that the specific terms actually play an important role in the model) except than θ_3 which is not significant and can be possibly eliminated from the model to obtain a simpler model. In the two left columns, the optimized parameters for the simpler ARMA(2,2) model are reported. Although the estimations were very close between the LSE and MLE estimators in the first case, here the estimations for MLE are different from both the LSE and the first more complex model.

6.6. Model Diagnostics

To test the goodness of fit for models and to investigate appropriate modifications over them we analyze the residuals and over-parameterization for these models. If the model is correctly specified with well-estimated parameters, the residuals would be approximately white noise with i.i.d. normal distribution having zero mean and constant variance. So by looking at the residuals' plot we do not expect to see non-zero mean or any trend in the series which is roughly the case for standardized residuals from our fitted ARMA(2,2) model in Figure 6.12. A point noticeable from this plot that still requires further analysis is the possible fluctuation of the variation of residuals which is critical in uncertainty modeling.

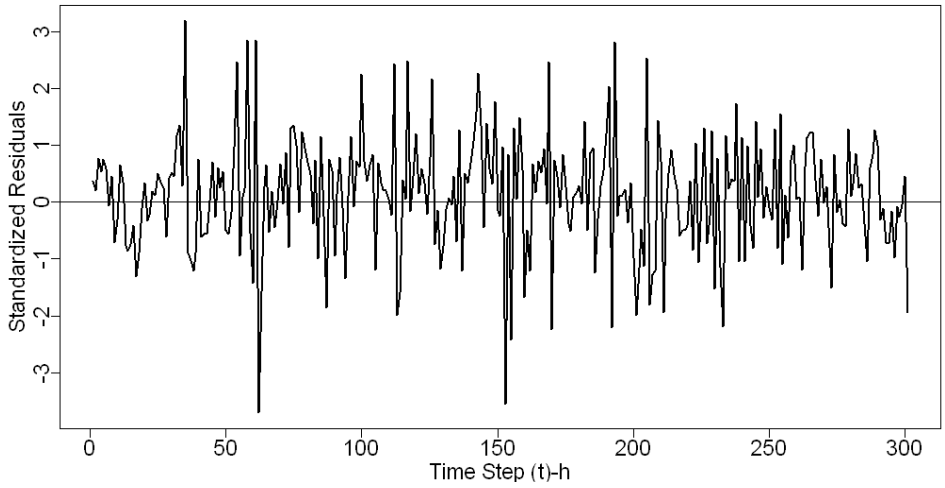


Figure 6.12. Standard residuals from the LSE ARMA(2,2) model

To assess the normality of the residuals the quantile-quantile plot of residuals are depicted in Figure 6.13. There is clear deviation from the theoretical normal quantiles in this plot. The Shapiro-Wilk normality test result also rejects the normality of residuals ($p<0.005$). The ACF of residuals can also provide valuable information about the independence of residuals. In Figure 6.14 the ACF of residuals from the temperature error ARMA(2,2) model is plotted. Although most lags confirm independence between residuals, the statistically significantly different from zero autocorrelations around lag 24 are very important. This information can lead us into proper further customization of our model for the series.

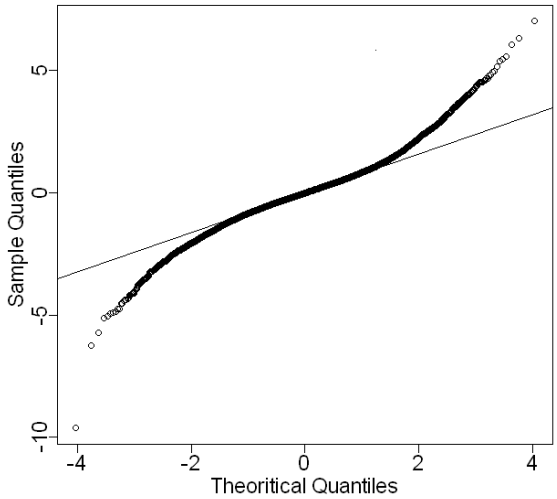


Figure 6.13. Quantile-quantile plot for residuals from the ARMA(2,2) model

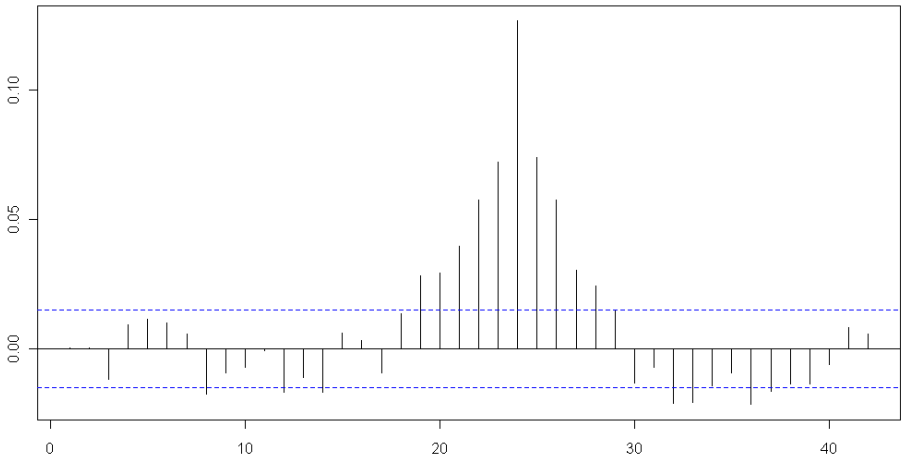


Figure 6.14. ACF of residuals from the ARMA(2,2) model

As another model specification technique, a close more general model is (over)fitted and the original model is accepted if the additionally introduced parameters are not significantly different from zero and the estimates of the original parameters are not dramatically changed. In Table 6.3 the basic ARMA(2,2) model is further investigated by trying close general forms. These forms are obtained by adding an order to either autoregressive or moving average components and hence considering ARMA(2,3) and ARMA(3,2). It can be noticed than in both cases the estimated parameter for the added order (θ_3 in the second column and ϕ_3 in the third column) is not significantly different from zero and hence the general model is rejected.

6.7. Forecasting

Using the available series up to time *t* (forecast origin) the forecast for Y_{t+l} occurring *l* (lead time) steps ahead is to be computed which is denoted as $\hat{Y}_t(l)$. It can be shown that to obtain the minimum mean square error we have [19][25]:

$$\hat{Y}_t(l) = E(Y_{t+l}|Y_1, Y_2, \dots, Y_t) \quad (6.32)$$

If there is a deterministic trend model involved in the process, it can be used to calculate the trend element in the lead time i.e. μ_{t+l}.

It can be shown that $Y_t(l) \approx \mu$ for large l in all stationary ARMA models. This is intuitive as the dependence between the forecast and observations gradually disappear until there is no information to improve on the naïve forecast of μ. Also for the variance of error we have $Var\big(e_t(1)\big) = \sigma_e^2$ and by using the MA(∞) form:

$$e_t(l) = e_{t+l} + \Psi_1 e_{t+l-1} + \Psi_2 e_{t+l-2} + \cdots + \Psi_{l-1} e_{t+1} \quad (6.33)$$

which holds for all ARIMA models. The forecast is unbiased i.e. $E\big(e_t(l)\big) = 0$ and:

$$Var\big(e_t(l)\big) = \sigma_e^2\big(1 + \Psi_1^2 + \Psi_2^2 + \cdots + \Psi_{l-1}^2\big) = \sigma_e^2 \sum_{j=0}^{l-1} \Psi_j^2 \quad (6.34)$$

meaning that with increasing lead time the error variance increases and $Var\big(e_t(l)\big) = Var(Y_t) = \gamma_0$ for large l. Please refer to [19] for details of obtaining this equation. It can also be shown that generally for nonstationary ARIMA processes the forecast error variance increases into the future for example in the random walk case we have $\Psi_j = 1$ for all *j* so $Var\big(e_t(l)\big) = l\sigma_e^2$. To further clarify the forecast process the explicit forecast expression for an ARMA(1,1) will be:

$$\hat{Y}_t(l) = \mu + \phi^l(Y_t - \mu) - \phi^{l-1} e_t \quad for\ l \geq 1 \quad (6.35)$$

Also in the forecasts of $l > q$ the autoregressive portion of the forecast equation for this process remains only, as we have $E(e_{t+l}|Y_1, \dots, Y_t) = 0$ for *j*>0.

Assuming an i.i.d normally distributed white noise terms for $\{e_t\}$ an ARIMA series will also have a normal distribution for $e_t(l)$ and hence the prediction limits for forecasts can be obtained by:

$$P\left[\hat{Y}_t(l) - z_{1-\alpha/2}\sqrt{Var\big(e_t(l)\big)} < Y_{t+l} < \hat{Y}_t(l) + z_{1-\alpha/2}\sqrt{Var\big(e_t(l)\big)}\right] = 1 - \alpha \quad (6.36)$$

where $z_{1-\alpha/2}$ is the critical value of standard normal distribution for being $(1-\alpha)100\%$ confident that the observation at step $t + l$ will be within the prediction interval. Note that $Var\big(e_t(l)\big)$ is obtained by Equation (6.34).

6.8. Improved Models

6.8.1. Daily Cycle

In many cases when there is clear autocorrelation at seasonal lags such as previous day, week, month, etc. we need to incorporate such correlation in the process. As a first step seasonal difference of period s is an appropriate transformation for modeling nonstationary seasonal processes:

$$\nabla_s Y_t = Y_t - Y_{t-s} \quad (6.37)$$

where for example for an hourly series with $s = 24$ the transformed series will represent changes from the previous day in successive hours. Multiplicative Seasonal ARIMA models are the general form of seasonal processes where for $\{Y_t\}$ an ARIMA$(p, d, q) \times (P, D, Q)_s$ is considered for which:

$$W_t = \nabla^d \nabla_s^D Y_t \quad (6.38)$$

is an ARMA$(p, q) \times (P, Q)_s$ model. In addition to autoregressive terms up to p lags and moving average terms up to q lags this model also includes P seasonal autoregressive and Q seasonal moving average terms with seasonal lag of s. These seasonal terms are $\Phi_1 Y_{t-s}, \Phi_2 Y_{t-2s}, \dots, \Phi_P Y_{t-Ps}$ and $\Theta_1 e_{t-s}, \Theta_2 e_{t-2s}, \dots, \Theta_Q e_{t-Qs}$, respectively.

Referring back to Figure 6.14 the seasonal autocorrelation is noticeable at lags 24. This is also confirmed by the BIC analysis as depicted in Figure 6.11. As for the possible differencing transformations here we apply the first difference and the seasonal difference ($D = 24$). Using these two transformations over the forecast error series the resulting time series plots (for the same window shown in Figure 6.3) are shown in Figure 6.15 and Figure 6.16, respectively.

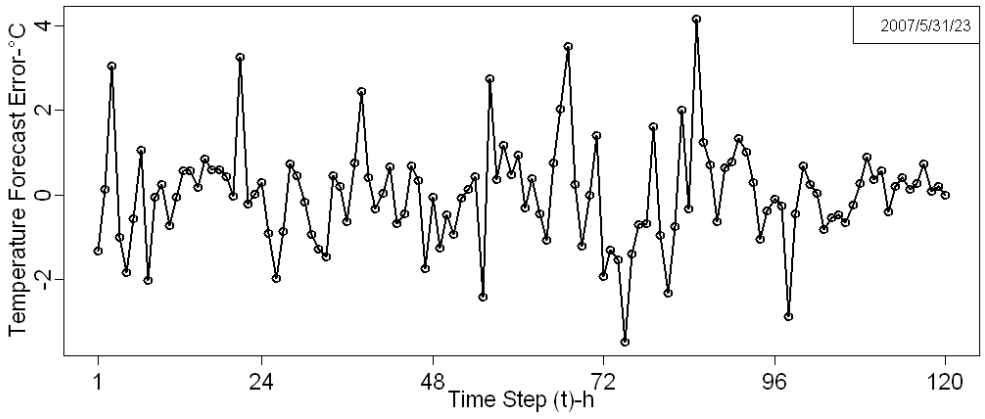


Figure 6.15. Transformed time series of temperature error using first difference ($d = 1$)

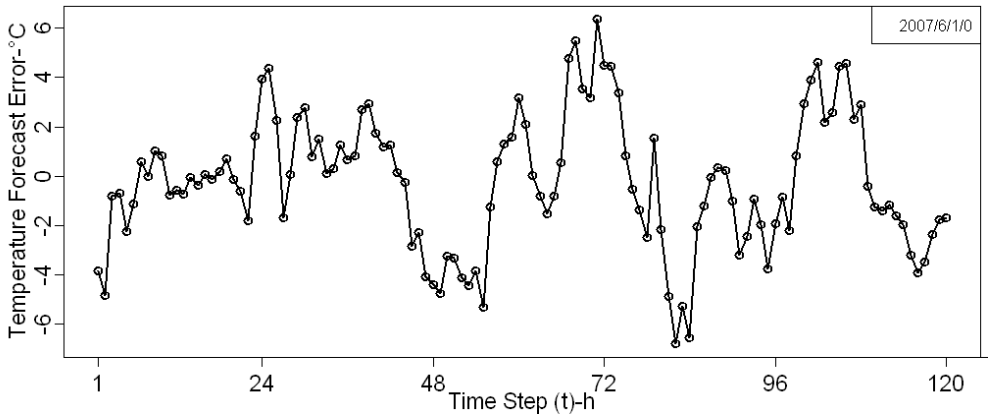


Figure 6.16. Transformed time series of temperature error using first seasonal difference ($D = 24$)

The choice between different transformations is made based on the seasonality removal ability of the transformation and its power to obtain a simpler yet accurate model for the series.

6.8.2. Cross-correlation

It is often the case that the target time series under analysis is related or impacted by other covariate series. Better models of forecast are expected to be obtained once such relevant covariates are incorporated. Assuming $X = \{X_t\}$ as a covariate time series for $Y = \{Y_t\}$, the cross-correlation function (CCF) between *X* and *Y* at lag *k* is defined as:

$$\rho_k(X,Y) = Corr(X_t, Y_{t-k}) = Corr(X_{t+k}, Y_t) \quad (6.39)$$

where *X* and *Y* are jointly (weakly) stationary if their means are constant and their cross-covariance $\gamma_{t,s}(X,Y)$ is a function of $t-s$. The sample cross-correlation function $r_k(X,Y)$ can be used to empirically investigate the lags at which a covariate series has influence on the target. The critical value for significantly different from zero sample cross-correlations magnitudes is $1.96/\sqrt{n}$ based on the assumption that *X* is independent of *Y* and hence $N(0,1/n)$ is the distribution for $r_k(X,Y)$. However, due to the autocorrelations present in *X* and *Y* this variance turns out to be inaccurate. In the case of stationary *X* and *Y* independent series [19]:

$$var\big(r_k(X,Y)\big) = \frac{1}{n}[1 + 2\sum_{k=1}^{\infty}\rho_k(X)\rho_k(Y)] \quad (6.40)$$

which can be much larger than 1/n with autocorrelations in *X* and *Y*. For non-stationary data the sample distribution will not be even normal. Hence, spurious cross-correlation can be easily detected even between independent series. It can be noted that the variance is $1/n$ if *X* and/or *Y* is a white noise process. This can be achieved by replacing the series

by the residuals from a fitted ARIMA model in practice. An AR(∞) representation can be used if X_t follows an invertible ARIMA(p, d, q) model:

$$\tilde{X}_t = (1 - \pi_1 B - \pi_2 B^2 - \cdots) X_t = \pi(B) X_t \quad (6.41)$$

where $\pi(B)$ is the filter which uses π_i parameters and the backshift operator (B) to obtain $\tilde{X}_t$ which is the residual at time t and hence white noise. This process is known as prewhitening. We can also obtain $\tilde{Y}_t$ using the same filter used for X and then calculate the CCF of $\tilde{X}$ and $\tilde{Y}$. As prewhitening is a linear operation, the original relationships will remain intact. The statistical significance of the sample CCF of the resulting prewhitened series can now be evaluated using the $1.96/\sqrt{n}$ threshold.

Thus a general linear regression model will be:

$$Y_t = \sum_{j=m_1}^{m_2} \beta_j X_{t-j} + Z_t \quad (6.42)$$

where X is independent of Z and Z_t can be modeled using ARIMA models. The lags of X present in the model can be determined by cross-correlation analysis.

In the context of modeling the temperature forecast error it seems promising to include relevant exogenous variables from parallel influential series such as the forecasted temperature (t2) and surface pressure (psf). Sample cross-correlation analysis results are provided in Figure 6.17 and Figure 6.18 which both confirm significant correlations in the zero lag. Also both series exhibit marginally significant cross-correlation around the diurnal lag. In addition the temperature series has noticeable cross-correlation in lag 1 too.

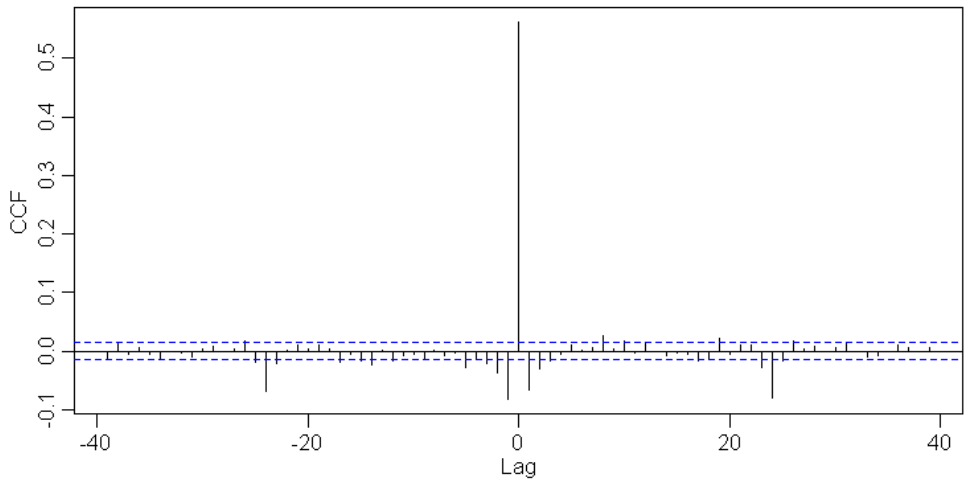


Figure 6.17. CCF plot of error series and the t2 forecast series

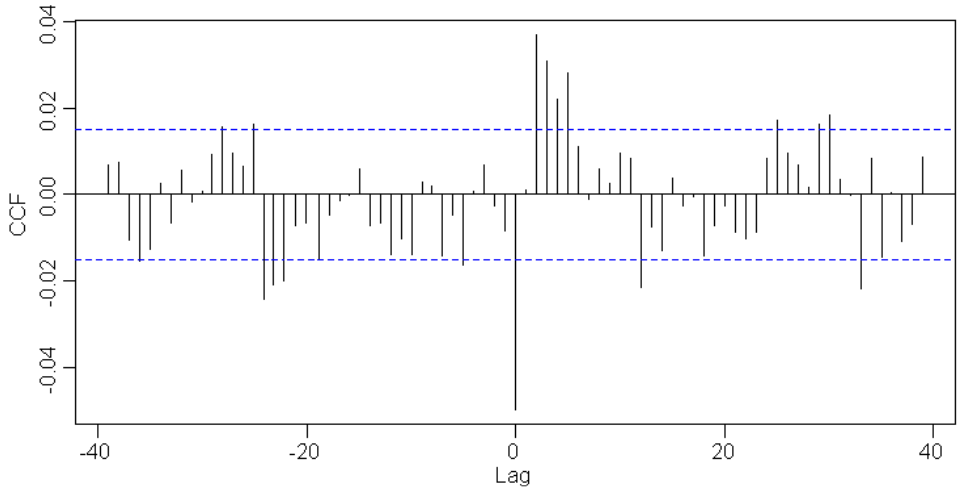


Figure 6.18. CCF plot of error series and the surface pressure forecast series

After performing similar analysis on the other available exogenous series (Table 4.1), the sixth lag of the relative humidity series (rh2) is also added to the ARIMA model along with the terms detected above from pressure and temperature. Sample window of the original forecast error series (Y_t) along with the exogenous linear regression series and its corresponding residual series (Z_t) are depicted for two cases of using the BF2 feature set only and using the BF2 feature set along with lagged features having significant cross-correlation with the target series in Figure 6.19 and Figure 6.20, respectively.

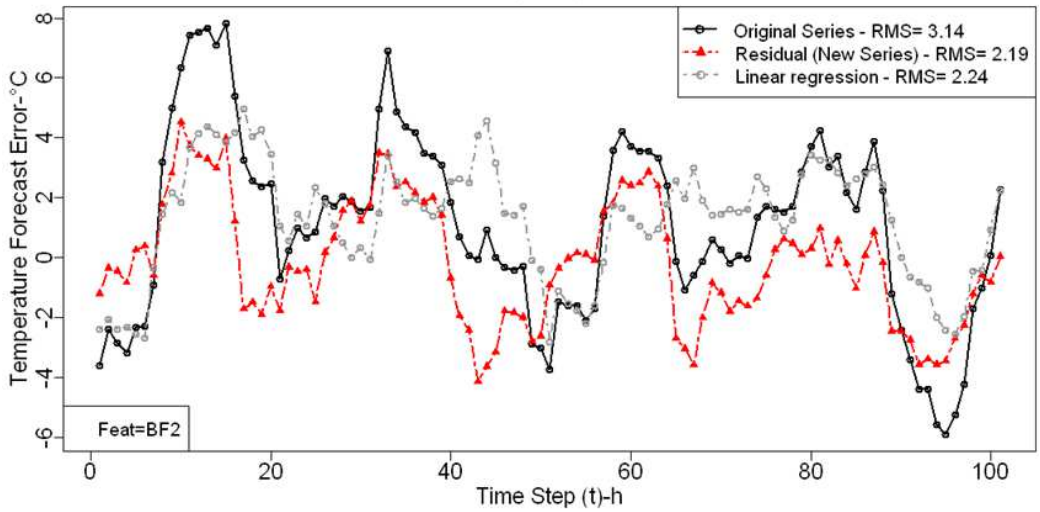


Figure 6.19. Residual and regression series using zero lag exogenous features of BF2

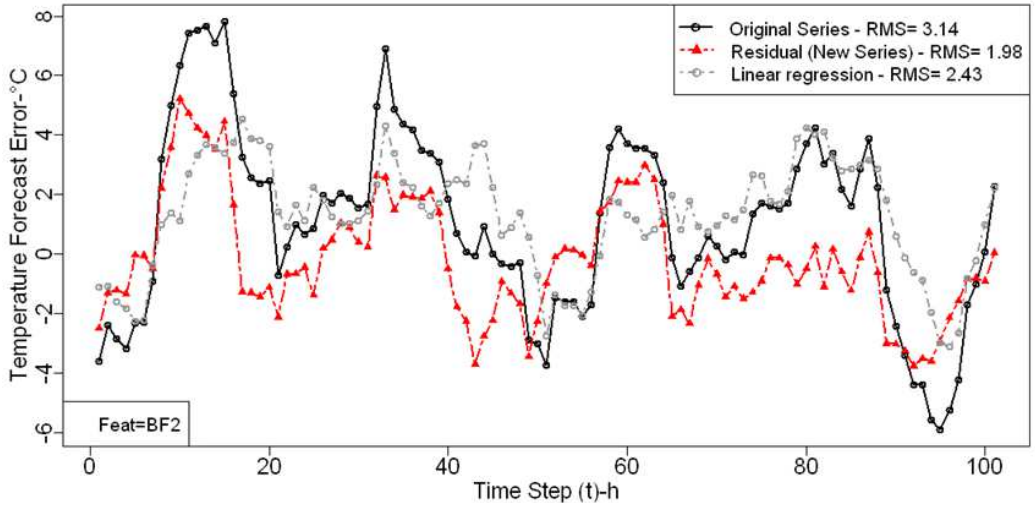


Figure 6.20. Residual and regression series using zero lag exogenous features of BF2 along with lagged features of t2, rh2 and psf

The RMSE of the newly obtained residual series in both cases confirm higher accuracy of the model when exogenous variables are accounted for. In addition results shown in Figure 6.20 confirm that the residual accuracy will improve to 1.98 from 2.19 when lagged variables are included in the model. Also note that the original time series is transformed using the seasonal mean model described in Table 6.1 as it was determined to be a better seasonal model in these experiments. All of these specification and diagnostics analysis results are in-sample and hence only use first two years of data with 80% proportion for training and 20% for validation. More related experimental results are provided in subsection 6.10.

6.9. Heteroscedasticity Modeling of Forecast Error

Rather than modeling the conditional mean of a time series as performed in ARIMA, there is an increasing interest in modeling the conditional variance of the series as an uncertainty measure. Instead of assuming a constantly increasing variance for forecasts of any number of steps ahead in ARIMA (refer to Equation (6.34)), the conditional variance can be considered as a random process by itself and hence modeled in connection to the current and past values (Homoscedasticity vs. Heteroscedasticity). For instance in many financial series, periods of larger volatility are often followed by larger conditional variance as opposed to stable periods [21].

Suppose we have noticed that recent temperature forecast errors have been unusually volatile. We might expect that the next hour's forecast error is also more variable than the typical volatility. However, an ARMA model cannot capture this type of behavior

because its conditional variance is constant. So we need other time series models in order to model the non-constant volatility.

Here we focus on the study of such dynamical patterns in the volatility of the temperature forecast error time series. The ACF, PACF and EACF results may show little and insignificant serial correlation in the ARIMA residual series, suggesting a *white noise* model. The sample ACF and PACF functions of the residuals from the best fitted seasonal exogenous ARIMA model (details in subsection 6.10) are plotted in Figure 6.21 and Figure 6.22. Both these plots suggest an i.i.d residual (considering the fact that few significant correlations can happen by chance). Similarly EACF results (not shown here) confirm a white noise model for the residuals.

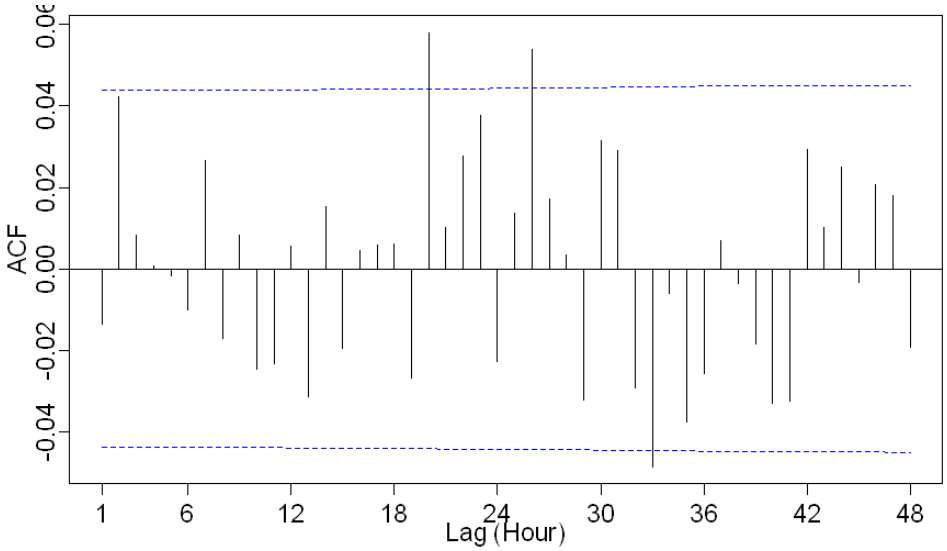


Figure 6.21. Sample ACF function of the residuals from the best ARIMA model

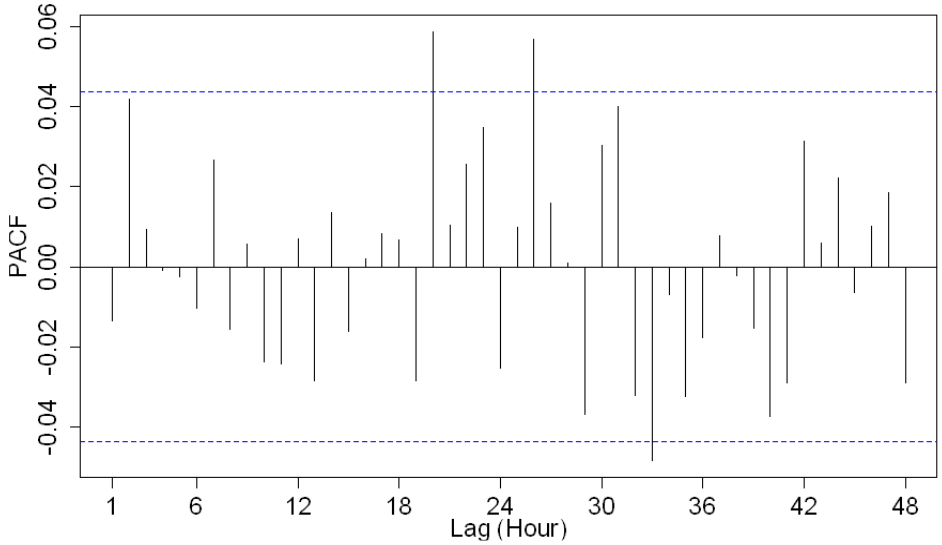


Figure 6.22. Sample PACF function of the residuals from the best ARIMA model

With an i.i.d. random variable for the residual series, transformations such as logarithms, absolute values or squaring must preserve independence. If otherwise there is some significant autocorrelations detected in the absolute or squared transformations of the original series, one can conclude on the existence of some higher-order dependence. Such dependence is evident in the ACF and PACF plots for absolute value of residual series in Figure 6.23 and Figure 6.24. Similar correlation was observed in the squared residual series (not shown here).

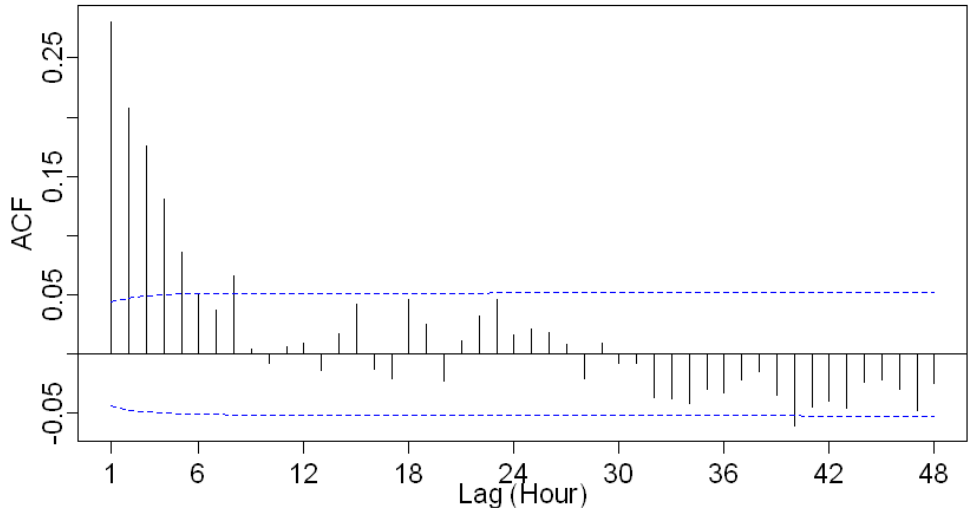


Figure 6.23. Sample ACF function of the absolute residuals from the best ARIMA model

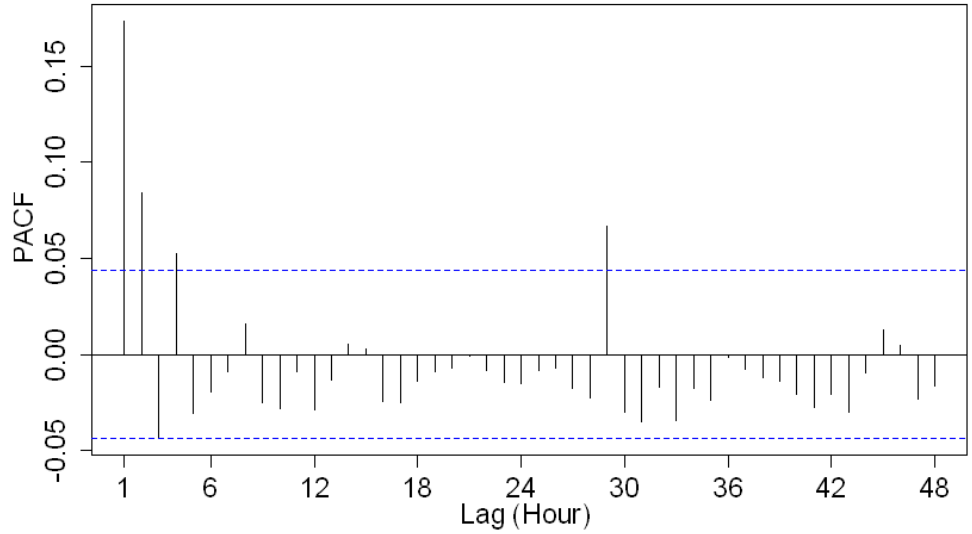


Figure 6.24. Sample PACF function of the absolute residuals from the best ARIMA model

In addition to these visual tools, the Box-Ljung test is often used to test autocorrelation in a series. Under the assumption that there is no AutoRegressive Conditional Heteroscedasticity (ARCH) present for the residuals of an ARMA model, the Box-Ljung statistic will have a chi-square distribution with *m* degrees of freedom for the

first *m* autocorrelations of the squared residual series and is called as the McLeod-Li test [50]. Conforming to the autocorrelation analysis the results from this test are all significant at the 5% significance level as shown in Figure 6.25.

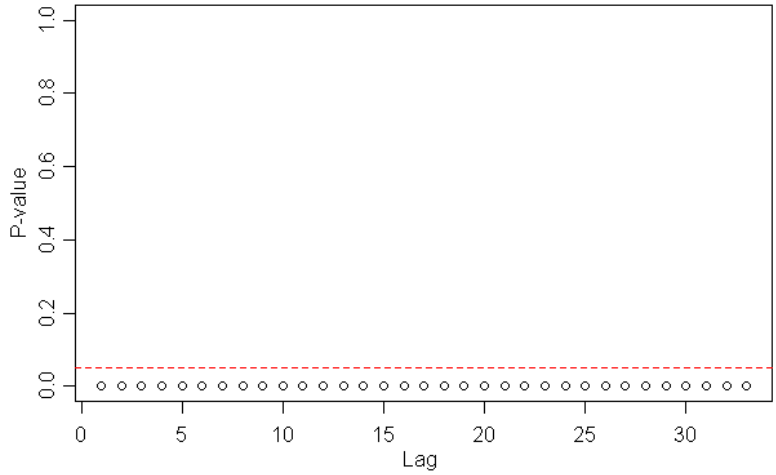


Figure 6.25. McLeod-Li test statistics for ARIMA residuals

All of these results show that although the series is serially uncorrelated but it has a higher-order dependence structure which is expressed as patterns of conditional variance (namely volatility clustering). Here we try to capture such patterns in a series using Generalized AutoRegressive Conditional Heteroscedasticity (GARCH) models.

AutoRegressive Conditional Heteroscedasticity (ARCH) proposed by Engle [21] model the variance of a time series. Given the original series $\{r_t\}$ up to time *t*-1, the conditional variance (or conditional volatility) of r_t is defined as $\sigma^2_{t|t-1}$. Based on the fact that r_t^2 is an unbiased estimation of $\sigma^2_{t|t-1}$ one can hypothesize that a period of large squared values can foretell a period with large variance and on the other hand a period of small squared values can foretell a stable period. The ARCH(1) model is a regression model with the conditional variance as its target variable using lag one of the squared values as its feature [58]:

$$r_t = \sigma_{t|t-1}\varepsilon_t \quad (6.43)$$

$$\sigma^2_{t|t-1} = \omega + \alpha r^2_{t-1} \quad (6.44)$$

where ω and α are unknown parameters and ε_t is i.id.d with zero mean and unit variance. To replace the conditional variance by some observable value the following definition is used:

$$\eta_t = r_t^2 - \sigma^2_{t|t-1} \quad (6.45)$$

where $\{\eta_t\}$ is a serially uncorrelated series with zero mean and is uncorrelated with past values of the original series. Using this equation in Equation (6.44):

$$r_t^2 = \omega + \alpha r_{t-1}^2 + \eta_t \quad (6.46)$$

Hence, under the assumption of having an ARCH(1) model for the original series, the squared series follows an AR(1) model. Based on the stationarity assumption of the $\{r_t\}$ series and taking expectations of the above equation:

$$\sigma^2 = \omega + \alpha\sigma^2 \quad (6.47)$$

where σ^2 is the stationary variance of the *r* series and is used for forecasting the *l* step ahead conditional variance. For *l*=1:

$$\sigma_{t+1|t}^2 = \omega + \alpha r_t^2 = (1-\alpha)\sigma^2 + \alpha r_t^2 \quad (6.48)$$

and generally $\sigma_{t+l|t}^2 = r_{t+l|t}^2$ for $l < 0$. As a more general approach the model can include q past squared terms of the series to obtain a ARCH(q) model. Also by adding p lags of the conditional variance to the model the Generalized AutoRegressive Conditional Heteroscedasticity model is defined as GARCH(p,q) [5][30]:

$$\sigma_{t|t-1}^2 = \omega + \beta_1\sigma_{t-1|t-2}^2 + \beta_2\sigma_{t-2|t-1}^2 + \cdots + \beta_p\sigma_{t-p|t-p-1}^2 + \alpha_1 r_{t-1}^2 + \alpha_2 r_{t-2}^2 + \cdots + \alpha_q r_{t-q}^2 \quad (6.49)$$

With nonnegative coefficient in a GARCH model the conditional variances are guaranteed to be nonnegative. Yet, this constraint is not essential for obtaining positive variances from the GARCH model. Using Equation (6.45) in the above definition of GARCH:

$$r_t^2 = \omega + (\beta_1 + \alpha_1)r_{t-1}^2 + \cdots + \left(\beta_{max(p,q)} + \alpha_{max(p,q)}\right)r_{t-max(p,q)}^2 + \eta_t - \beta_1\eta_{t-1} - \cdots - \beta_p\eta_{t-p} \quad (6.50)$$

where $\beta_k = 0$ for $k > p$ and $\alpha_k = 0$ for all $k > q$. Hence, for $\{r_t\}$ series following the GARCH(p,q) model, the $\{r_t^2\}$ series is an ARMA(max(p,q), p). Yet due to the larger sampling variability for higher moments, the order analysis is usually done using the absolute series i.e. $\{|r_t|\}$. Details of stationarity conditions and proofs are discussed in [19]. Finally the trained model can be used to forecast the *l*-step-ahead conditional variance. As an example using a GARCH(1,1) model and using $\sigma^2 = \alpha\sigma^2 + \omega/(1-\alpha_1-\beta_1)$:

$$\sigma_{t+1|t}^2 = (1-\alpha_1-\beta_1)\sigma^2 + \alpha_1 r_t^2 + \beta_1\sigma_{t|t-1}^2 \quad (6.51)$$

thus the next conditional variance is a weighted average of the long-run variance, the last available squared observation and the last prediction of the variance. By assuming normal innovations, the likelihood function of a GARCH model can be obtained and then numerically optimized to estimate the model coefficients and their corresponding variance.

It is evident that the GARCH model can only capture the conditional variance of the process under analysis. Hence, in order to model the conditional mean of the $\{Y_t\}$ series the ARMA model is still needed to be utilized. In this setting the GARCH(p,q) model is used to model the white noise term in the conditional mean's ARMA(a,b) model:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_a Y_{t-a} + e_t - \theta_1 e_{t-1} - \cdots - \theta_b e_{t-b} \quad (6.52)$$

$$e_t = \sigma_{t|t-1}\varepsilon_t \quad (6.53)$$

$$\sigma^2_{t|t-1} = \omega + \beta_1 \sigma^2_{t-1|t-2} + \cdots + \beta_p \sigma^2_{t-p|t-p-1} + \alpha_1 e^2_{t-1} + \cdots + \alpha_q e^2_{t-q} \quad (6.54)$$

where the ARMA and GARCH orders can be determined by analyzing the $\{Y_t\}$ and $\{e_t^2\}$ series, respectively. The parameters can then be estimated independently. After model diagnosis confirms the two models, they can be applied to forecast both the conditional mean and variance of target future values.

The extended autocorrelation sample function of the absolute (or squared) residuals can provide an analysis on the order of the GARCH model. Results for the absolute residual series of the temperature forecast error ARIMA model in Table 6.4 suggest a GARCH(1,1) model.

Table 6.4. Sample EACF for absolute residuals of the best ARIMA model

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	o	o	o	o	o	o	o	x	x
1	x	**o**	o	o	o	o	o	o	o	o	o	o	x	o
2	x	x	o	o	x	o	o	o	o	o	o	o	o	x
3	x	x	o	o	o	o	o	o	o	o	o	o	o	x
4	x	x	x	x	o	o	o	o	o	o	o	o	o	o
5	x	x	x	o	x	o	o	o	o	o	o	o	o	o
6	x	x	o	o	x	x	o	o	o	o	o	o	o	o
7	x	x	x	x	o	o	x	x	o	o	o	o	o	o

To accept a fitted model a major assumption to verify is whether the standard residuals i.e. $\hat{\varepsilon}_t = r_t / \hat{\sigma}_{(t|t-1)}$ are independently and identically distributed. Looking at the ACF of the absolute or squared standard residuals one can check for serial autocorrelation and volatility clustering. Figure 6.26 shows the ACF of standardized

residuals from the GARCH(1,1) model which was found to be the best GARCH model for the residuals of the best ARIMA model fitted to the temperature error series. This plot has the general impression that the residuals are no longer serially correlated and hence the volatility clustering has been well captured in the model.

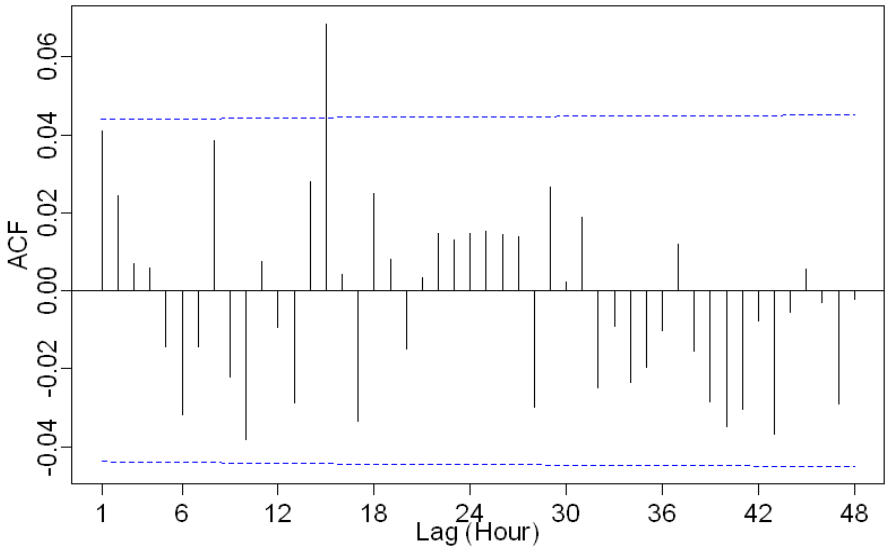


Figure 6.26. Sample ACF of absolute standard residuals from the fitted GARCH(1,1) model

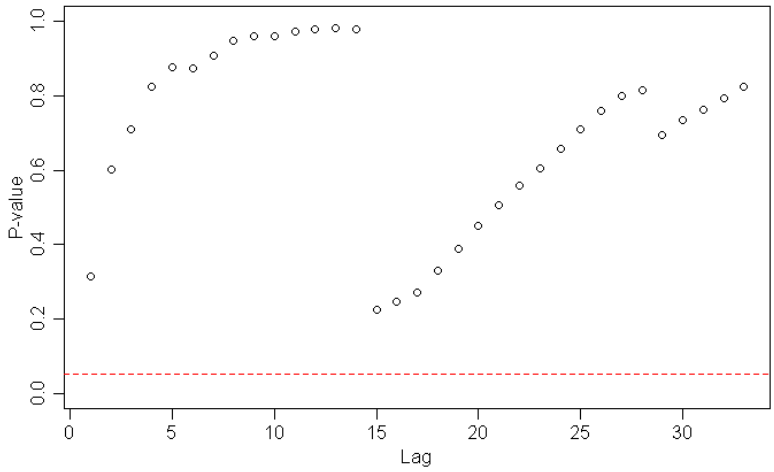


Figure 6.27. McLeod-Li test statistics for GARCH(1,1) residuals

The results from the McLeod-Li test also confirm that the residuals do not exhibit serial autocorrelation anymore and hence the fitted GARCH model is a good candidate for the conditional variance of the original series.

6.10. Experimental Results

6.10.1. Data Sets and Method Set-ups

After performing model specification analytics in the previous subsection, here we focus on the practical application of time series models for the temperature forecast error and uncertainty modeling and prediction. We use the first two years of data (i.e. 2007 and 2008) from the AG data set for training and diagnostics and preserve the 2009 data for out-of-sample forecasting and evaluation in each of the two stations. Different time series models are compared to the clustering and quantile regression methods in terms of both their point and interval forecast accuracy and skill. We also evaluate the time series models relative to three baseline competitors. The first baseline method is the no-change forecast known as the "persistence forecast" in the climatological literature (referred to as Persistence). The two other baselines are the best simple moving average models using the past 4 hours (MA-H4) and the same hour values in the past three days (MA-D3), respectively.

For prediction interval forecasts of the ARIMA time series models we use the theoretical variance approach with a Gaussian assumption and the exponential smoothing multi-step-ahead error variance formula (Equation (6.34)). For the baseline models the empirical distribution of the forecast errors (in a 1000 hour window) for each lead time is used to obtain variance estimations. Finally, the GARCH model can provide a series of conditional variance forecasts (and hence prediction intervals) under the Gaussian and empirical distribution assumption for any lead time. The results from any of the two distribution assumptions were very close and in favor of the empirical distribution in rare cases. Hence, we only provide the results from the empirical distribution assumption here. Also Analytical results did not support different model specifications for different seasons and stations. However, independent models (with identical specification) were fit for each station.

6.10.2. Forecast Evaluation Results

We first consider the performance of the various time series models in terms of point (i.e. expected mean value) forecasts. The RMSE of point forecasts made for different forecast horizons in the two stations are provided in Table 6.5. For ARIMA models there are three entries ARIMA, sARIMA and sxARIMA marking the three different setups of: simple ARIMA ($p = 2$, $d = 1$, $q = 2$), Seasonal ARIMA ($p = 2$, $d = 1$, $q = 2$, $P = 2$, $D = 0$, $Q = 1$) and Seasonal Exogenous ARIMA (including BF2 and lagged features as

described in Figure 6.20). Optimization methods of LSE and MLE yield very similar results with marginally better performance by LSE models. The results clearly show the higher accuracy of the sxARIMA model in these point forecasts.

The average RMSE in both stations for these models are plotted in Figure 6.28. Considering the forecasting performance of the baseline models in the first horizons, it can be seen that the Persistence has higher accuracy. However, with increasing horizon length the Persistence forecast becomes comparatively poor. Yet, the MA-D3 model has a better accuracy in longer leads and is more persistent in its accuracy. As expected, the sxARIMA model consistently outperforms the other models in terms of forecast accuracy and therefore is used for the prediction interval modeling and forecasting and referred to as ARIMA here after.

Table 6.5. Point forecast accuracy of time series models in terms of RMSE

	1-h	**2-h**	**3-h**	**6-h**	**12-h**	**24-h**	**48-h**
Agassiz							
Persistence	1.10	1.73	2.27	3.43	4.45	2.64	2.75
MA-H5	2.22	2.50	2.79	3.64	4.18	2.64	2.96
MA-D3	1.82	2.19	2.26	2.34	2.38	2.47	2.55
ARIMA	1.00	1.48	1.86	2.41	2.59	2.61	2.76
sARIMA	0.98	1.41	1.71	2.18	2.38	2.28	2.52
sxARIMA	0.81	1.14	1.41	1.79	1.98	1.98	2.16
Hope							
Persistence	1.08	1.79	2.31	3.33	4.05	3.02	3.23
MA-H4	2.10	2.44	2.73	3.42	3.84	2.96	2.98
MA-D3	2.06	2.47	2.62	2.70	2.66	2.71	2.87
ARIMA	1.02	1.64	2.06	2.67	2.89	2.83	2.82
sARIMA	1.00	1.59	1.97	2.50	2.66	2.64	2.74
sxARIMA	0.75	1.28	1.59	1.99	2.04	2.04	2.18

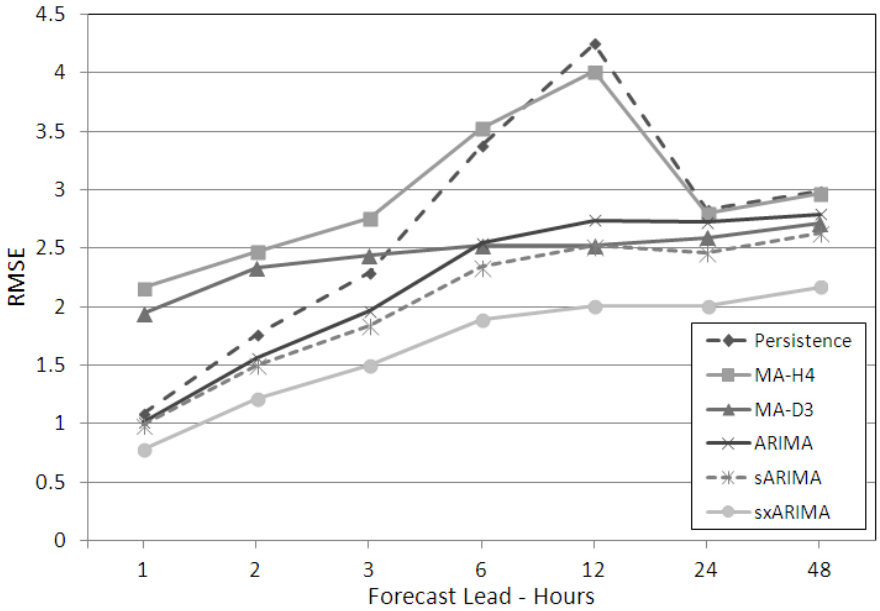


Figure 6.28. Average of point forecast accuracy in both stations for different time series models

Sample point and prediction interval forecasts from the ARIMA model for the Hope station in 2009 are plotted in Figure 6.29. As can be seen, the forecast series can catch the general shape of the series and matches well with the observations during the first leads. After longer leads (e.g. 40 hours) the forecasts start to smooth as there is less information available for the forecast. The prediction intervals also reflect on the increasing uncertainty in the forecasts as the forecast horizon increases. This is more evident in a closer look into the prediction interval forecasts of GARCH shown in Figure 6.30.

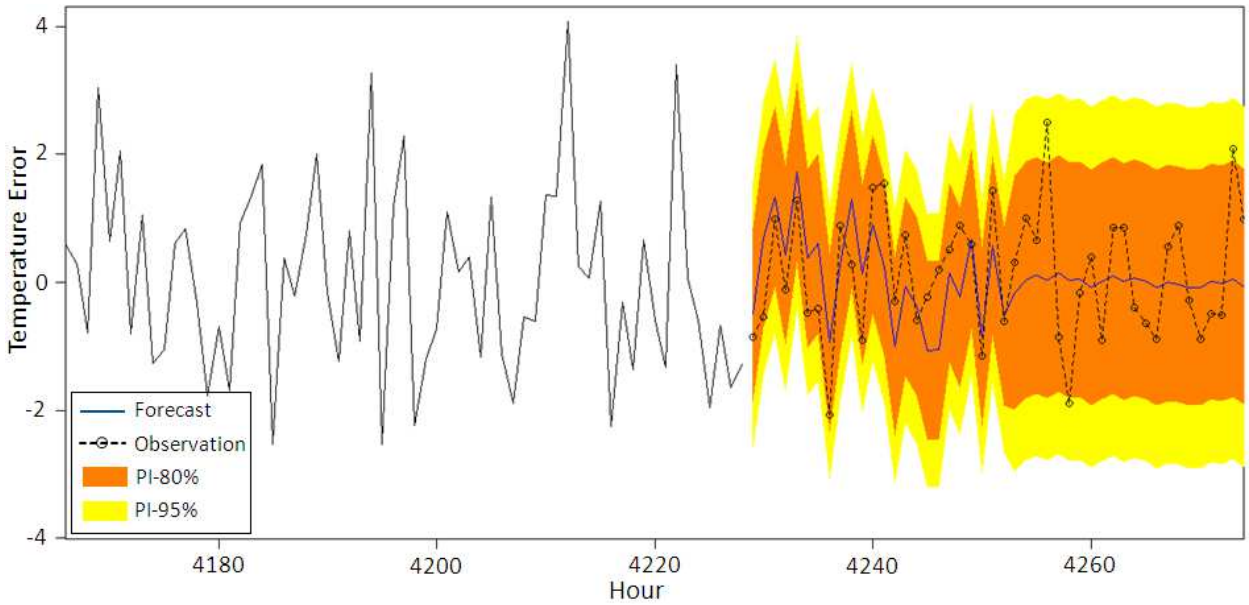


Figure 6.29. Sample ARIMA forecast along with theoretical prediction intervals

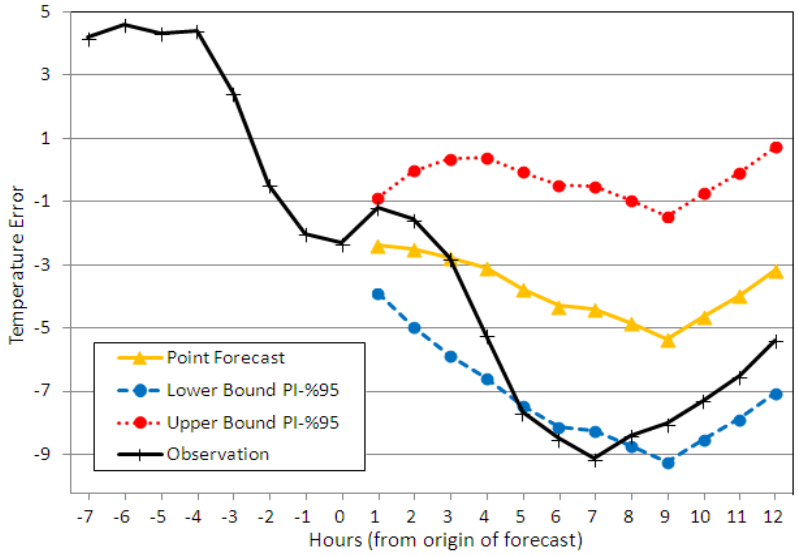


Figure 6.30. Sample GARCH prediction interval forecasts

Detailed evaluation measures for the prediction interval forecasts of the ARIMA model are listed in Table 6.6. Rather than having a single row in this table like the previous methods of clustering and quantile regression (which do not have any temporal awareness), the ARIMA model has an independent row for each lead time (in the forth column) as this has a critical role in the accuracy of the time series forecasts. As mentioned before this best model incorporates the BF2 and the lagged features (BF2+L) as regression inputs into the ARIMA time series model. The ARIMA forecasts outperform the best quantile regression model (i.e. SPQR) up to the 6-hour lead in terms of skill ($SScore^{0.95}$). The time series PI forecasts have comparable performance with quantile regression models up to 24-hour-ahead. After this lead, the ARIMA model is less accurate than quantile regression but yet considerably better than FCM and baseline models. The ARIMA model has a better coverage measure and also has a wider prediction interval when compared to quantile regression forecasts except than the first few leads.

In the next step an ARIMA(2,1,2)X(1,0,1)-GARCH(1,1) model is fit to the series meaning that the residuals of the ARIMA model are modeled by a GARCH(1,1) independently. This model is referred to as GARCH here. The order of the GARCH model is determined by the analysis performed in subsection 6.9. It must be noted that the point forecasts of the GARCH model are identical to that of the ARIMA model. However, the prediction interval forecasts are provided independently for each origin by the GARCH model as opposed to ARIMA where for each lead the variance will be approximately constant from any origin. This can be easily noted in Figure 6.31 which

plots the next-hour forecast of standard deviation for the GARCH target which is the ARIMA residual.

Table 6.6. Prediction interval verification measures for top models of different methods for 2009

Algorithm	*K*	Features	Fit/ Params	Sharpness (°C)	Coverage %	Coverage$^{0.95}$ %	Resolution	RMSE	*SScore*	*SScore*$^{0.95}$
SPQR	(50)	BF2	df=4	6.77	92.78	90.19	1.79	2.02	0.2231	0.2450
LocQR	(50)	BF2	λ=0.7	7.04	92.62	90.42	1.74	2.08	0.2290	0.2505
NLQR	(50)	BF2	-	7.00	92.57	89.95	1.86	2.06	0.2319	0.2550
KQR	(50)	BF2	σ=0.0042 C=4	7.22	91.98	90.14	1.93	2.15	0.2521	0.2839
LQR	(50)	BF2PG	-	7.84	93.52	91.03	1.57	2.26	0.2497	0.2719
FCM	45	BF2	Kernel	10.44	93.28	90.91	1.52	2.90	0.3413	0.3697
Base-Month	12	Month	Kernel	11.90	93.43	92.26	1.86	3.27	0.3774	0.3916
Base-Temp.	10	Temp.	Normal	11.41	93.09	92.15	0.79	3.16	0.3730	0.3849
Base-Clim.	1	-	Normal	11.89	92.95	92.62	0	3.25	0.3947	0.3993
ARIMA	(50)	BF2+L	l=1	3.05	93.82	91.30	0.09	0.78	0.1149	0.1334
ARIMA	(50)	BF2+L	l =2	4.73	94.36	91.97	0.19	1.21	0.1622	0.1828
ARIMA	(50)	BF2+L	l =3	5.84	94.44	92.08	0.41	1.50	0.1930	0.2141
ARIMA	(50)	BF2+L	l =6	7.29	94.30	91.94	0.51	1.89	0.2325	0.2531
ARIMA	(50)	BF2+L	l =12	7.84	94.63	92.33	0.28	2.01	0.2478	0.2698
ARIMA	(50)	BF2+L	l =18	7.96	95.18	92.94	0.23	2.03	0.2472	0.2692
ARIMA	(50)	BF2+L	l =24	8.01	94.99	92.70	0.23	2.01	0.2493	0.2716
ARIMA	(50)	BF2+L	l =48	8.28	94.29	91.91	0.31	2.17	0.2632	0.2875
ARIMA	(50)	BF2+L	l =120	8.83	94.68	92.36	0.65	2.30	0.2746	0.2980
ARIMA	(50)	BF2+L	l =240	9.58	95.05	92.80	1.20	2.43	0.2942	0.3192

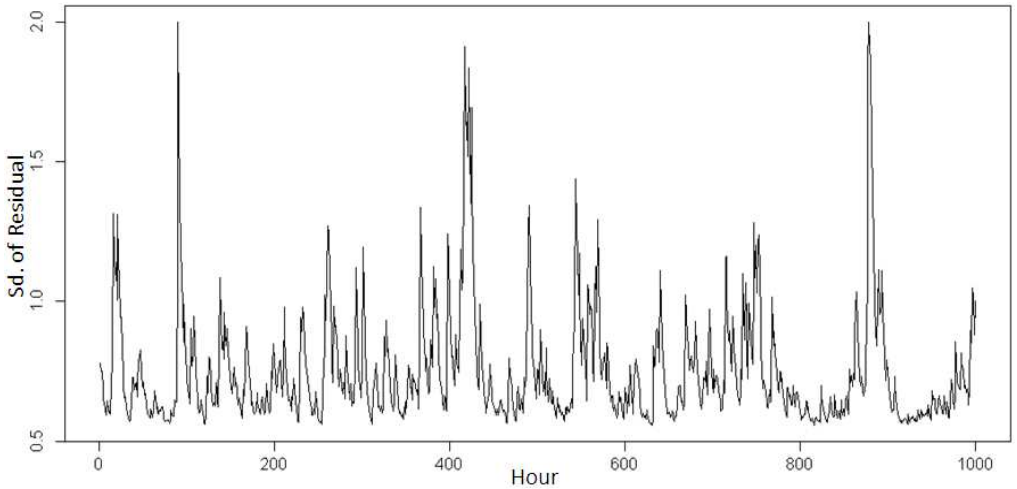


Figure 6.31. Sample forecast of next-hour sd. of error using GARCH(1,1)

A detailed comparison of uncertainty forecasts from MA-D3, ARIMA and GARCH in different leads are provided in Table 6.7. These results also confirm better performance of ARIMA and GARCH versus the baseline model. In addition prediction interval

forecasts from GARCH outperform ARIMA in the first three hours and are very comparable with it in the following leads into the future (Figure 6.32). The average width of the prediction intervals from the moving average model are clearly larger than ARIMA and GARCH, with GARCH having slightly wider intervals with higher coverage as shown in Figure 6.33 and Figure 6.34.

Table 6.7. Prediction interval verification measures for time series models

	1-h	2-h	3-h	6-h	12-h	24-h	48-h	120-h	240-h
Width									
MA-D3	7.80	9.51	9.89	10.00	10.00	10.03	10.73	11.97	12.59
ARIMA	3.05	4.73	5.84	7.29	7.84	8.01	8.28	8.83	9.58
GARCH	3.02	4.70	5.93	7.69	8.22	8.22	8.99	9.90	10.25
Coverage$^{0.95}$									
MA-D3	92.47	92.45	92.44	92.45	92.46	92.87	92.53	93.27	93.03
ARIMA	91.30	91.97	92.08	91.94	92.33	92.70	91.91	92.36	92.80
GARCH	91.97	91.57	91.83	92.42	93.24	93.48	93.98	94.65	94.61
Resolution									
MA-D3	0.72	0.92	0.97	0.99	0.99	1.00	1.10	1.55	1.45
ARIMA	0.09	0.19	0.41	0.51	0.28	0.23	0.31	0.65	0.61
GARCH	0.90	1.38	1.66	1.89	1.04	0.34	0.26	0.55	0.54
SScore$^{0.95}$									
MA-D3	0.2663	0.3247	0.3379	0.3416	0.3415	0.3464	0.3655	0.3976	0.4334
ARIMA	0.1334	0.1828	0.2141	0.2531	0.2698	0.2716	0.2875	0.2980	0.3192
GARCH	0.1240	0.1793	0.2130	0.2606	0.2694	0.2689	0.2847	0.2980	0.3151

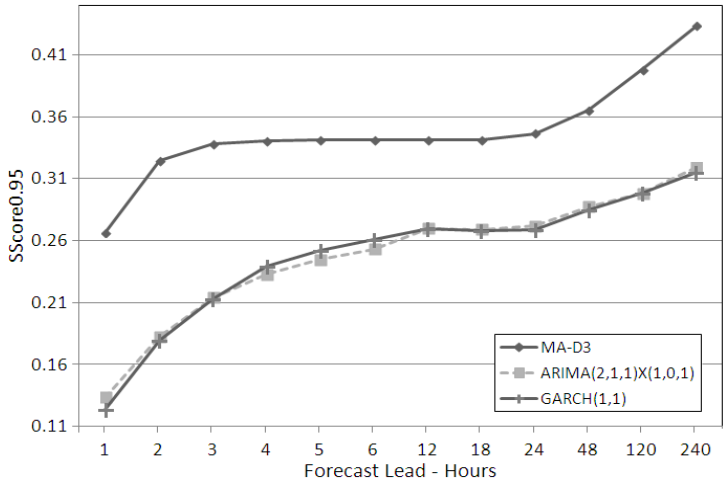


Figure 6.32. SScore95 over increasing forecast leads up to 10 days for different time series forecasting methods

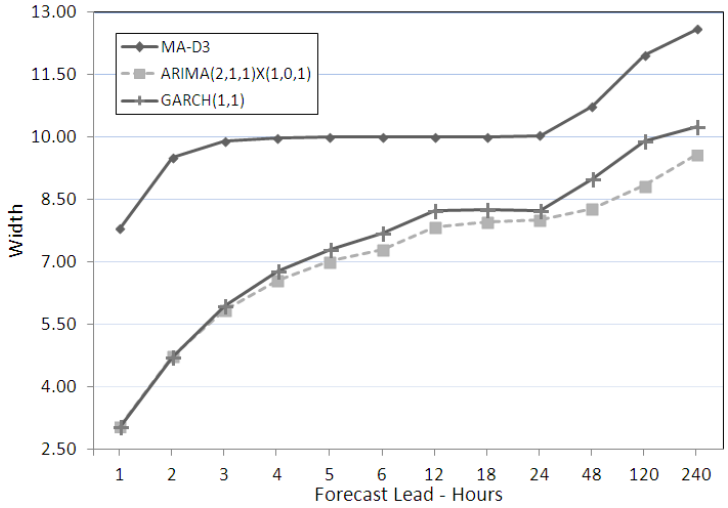


Figure 6.33. Time series forecast PI width comparison for up to 10 day-ahead forecast

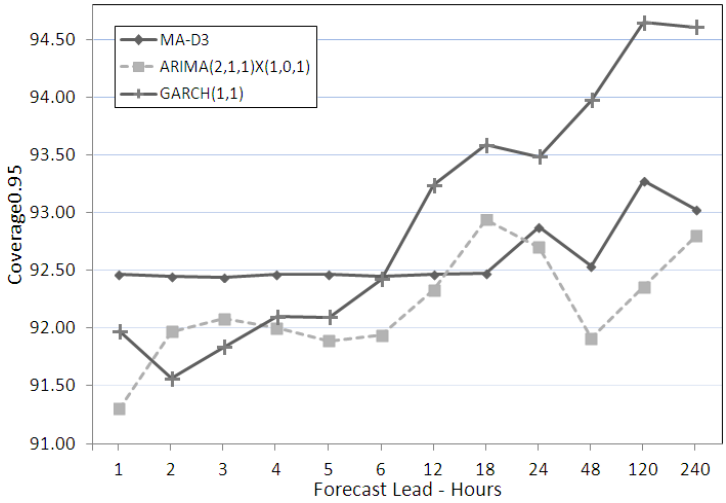


Figure 6.34. Time series forecast PI Coverage$^{0.95}$ comparison for up to 10 day-ahead forecast

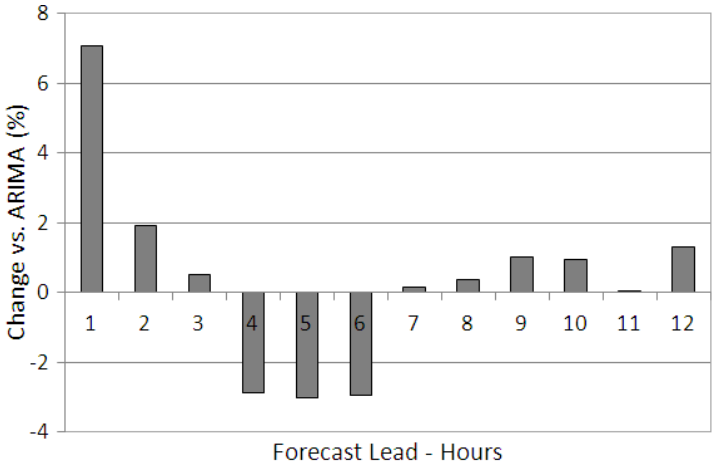


Figure 6.35. Percentage change of uncertainty forecast skill of GARCH compared to ARIMA

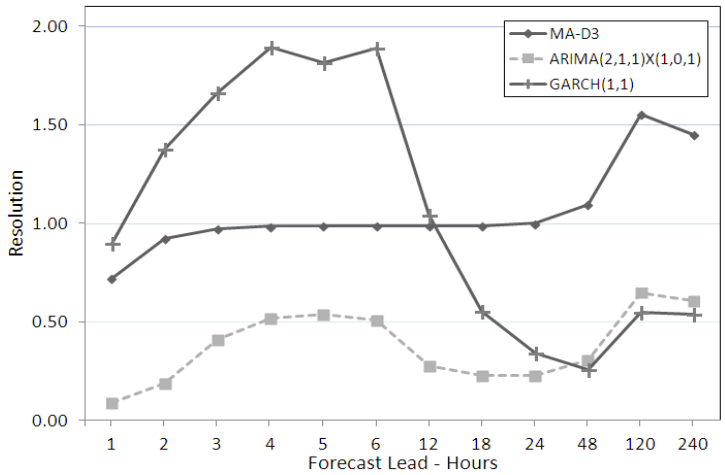


Figure 6.36. Time series forecast PI resolution comparison for up to 10 day-ahead forecast

To have a closer look at the skill of the prediction interval forecasts from GARCH Figure 6.35 plots the relative change of skill in GARCH compared to ARIMA in terms of percentage. The better skill of forecasts in the first leads is evident in this graph. There is also a lower accuracy of prediction interval forecasts observed in the 4, 5 and 6-hour-ahead forecasts. This is possibly due the weaker presence of the autoregressive structure of variance in further leads. However, by also considering the significantly higher resolution of the interval forecasts from the GARCH model (Figure 6.36) this model may be preferred for uncertainty modeling especially for smaller leads.

The empirical results confirm the advantage of the time series model in the application of weather forecast uncertainty modeling. In a practical set up the time series model can be utilized to provide better point and interval forecasts for up to 6 hours and SPQR can be employed for further leads to obtain the optimal result.

6.11. Conclusions

In this chapter, we looked into the application of time series models for the modeling of uncertainty. After basic time series analysis of the NWP temperature forecast error time series, different ARIMA models that incorporate seasonality, regression and cross-correlation were fitted to the series in two stations. Rather than using theoretical relations for conditional variance forecasting, heteroscedastic time series models (i.e. GARCH) were also studied as tools of modeling forecast uncertainty. These models explicitly focus on conditional variance as an independent time series model. In the experiments we applied these approaches along with some baseline time series models into the NWP forecasts and computed out-of-sample forecasts of point and prediction interval. The results clearly confirm the better skill of interval forecasts from the ARIMA and GARCH

models in the first few leads when compared to the best quantile regression model (i.e. SPQR). The point forecast accuracy of the ARIMA model outperforms that of SPQR for up to 1-day-ahead forecasts. Also, the GARCH model outperforms the ARIMA model in terms of uncertainty forecasts in the first three hour leads. This model also consistently provides prediction intervals with higher resolution. It can be generally concluded that in a practical set up the GARCH model and SPQR can be utilized simultaneously to provide forecasts in the first leads and longer leads, respectively.

As directions for future research, alternative versions of the GARCH model such as Asymmetric GARCH models can be investigated. Also, application and adaption of non-linear time series for the modeling of NWP forecast uncertainty is of interest.

Chapter 7
Conclusions and Future Directions

This chapter ends the thesis by providing concluding remarks about the various methods and experiments, the limitations of the study and open directions for enhancements and future research.

7.1. Summary and Discussion

In this thesis we studied the problem of modeling NWP forecast uncertainty by utilizing the point forecast accuracy records with a focus on learning methods. In this regard, we developed a comprehensive methodology to obtain accurate prediction interval forecasting models from NWP performance history. Firstly, clustering algorithms were proposed as an improvement over the classical approach of manual grouping of the forecast situations. Various clustering methods along with different distribution estimation methods were applied for this purpose.

Two different data sets of NWP forecasts and NCAR observations were used to practically test these uncertainty prediction models. A comprehensive evaluation framework was developed and used in the experiments that add the crucial aspect of sampling uncertainty in forecast skill measurements majorly absent in similar previous studies. The evaluation results clearly confirm the higher accuracy of the prediction interval forecasts obtained from these models (specifically the Fuzzy C-means model) when compared to the baseline and previously proposed methods. In addition, these models do not suffer from the dimensionality limitations of the previous methods.

In the next step, we further extended our methodology by implementation and evaluation of a wide range of quantile regression methods as alternative solutions to the prediction interval forecasting problem. Extensive experiments affirm the superior performance of some quantile regression methods (specifically Spline Quantile

Regression) when compared to the best clustering methods. It should be noted however, that the uncertainty models obtained using clustering methods have the advantage of directly providing the full density for the forecast target.

Time series modeling of the forecast error series was the focus of the last phase of this thesis. Various time series models including seasonal ARIMA models using exogenous elements were fitted to the forecast error series to obtain forecasts of prediction intervals and expected error into the future in increasing forecast horizons. Also, heteroscedastic models of forecast target variance which can model the conditional variance of NWP forecast error as an independent process were studied as alternative solutions to the problem of prediction interval modeling. Elaborate experimental studies proved that the best time series models (specifically combination of ARIMA and GARCH) can provide significantly more skilful prediction interval forecast for up to 6 hours ahead. In longer forecast horizons the time series model is outperformed by SPQR but yet can maintain a nearly comparable performance when compared to quantile regression and still outperforms the clustering and baseline models.

Based on this study, in a practical scenario a combination of the two best models can be employed, i.e. the ARIMA-GARCH time series model for the near future prediction interval forecasts (first 6 hours) and SPQR for further time steps into the future. It is also shown that these models are able to significantly improve the accuracy of the NWP point forecasts by incorporating a dynamic de-biasing process.

The research conducted in this thesis, leading to the proposed methodology, clearly confirms the feasibility and benefits of using point forecast performance databases to extend these forecasts into prediction intervals that can also provide critical information about the uncertainty of the forecasts. This uncertainty is modeled dynamically and is dependent on various influential aspects included in the model such as the current and recent past weather attributes (forecasted situations), the recent past accuracy of the system and station attributes such as elevation. Such information is of high value in various applications utilizing these forecasts in decision making and optimization process such as DTR and wind energy markets.

7.2. Future Directions

A limitation of this study was the unavailability of a sufficiently large NWP forecast performance database which is well geographically distributed to study spatial aspects and dependencies in uncertainty modeling. Moreover, various meteorological analytics

can be performed to provide new features (such as detection of troughs) that can potentially improve the accuracy of the uncertainty models. It will be also interesting to incorporate various ensemble-based uncertainty forecasts in the evaluation studies.

The algorithms and methods studied in this work can be extended in a few interesting directions:

(I) Improving the clustering prediction interval models by developing techniques to guide the clustering process (e.g. merging and splitting of clusters) using characteristics of forecast error distribution in the clusters.

(II) In the operational use of the PI forecasting methods, the model should be adaptive and update its parameters as the accuracy of new forecasts are revealed. Evolving Clustering Methods (ECM) [1] can be employed in this context.

(III) Developing proper kernel functions to optimally and efficiently learn non-linear quantile regression functions. Also using Artificial Neural Network (ANN) for faster learning of nonlinear quantile functions [11].

(IV) Currently, the quantiles of various confidence levels can happen to cross one another in the quantile regression models. Although, the frequency of such cases is very low, approaches imposing constraint on the learning problem can alleviate this problem in future study.

(V) Application of alternative versions of the GARCH model such as Asymmetric GARCH models [31]. Also, application and adaption of non-linear time series for the modeling of NWP forecast uncertainty is of interest.

(VI) Modeling the conditional density of the time series forecasts using models that consider a wider range of distribution moments and characteristics such as the method proposed in [30].

Bibliography

[1] Aggarwal, C.C., Han, J., Wang, J., Yu., P.S., 2003. A framework for clustering evolving data streams. In *Proceedings of the 29th Int Conf on Very Large Data Bases.*

[2] Anderson, J.L., 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*. 9(7), 1518–1530.

[3] Bessa, R.J., Miranda, V., Botterud, A., Wang, J., Constantinescu, E.M., 2012. Time adaptive conditional kernel density estimation for wind power forecasting. *IEEE Tran. on Sustainable Energy*. 3(4), 660-669.

[4] Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences*. 10(2), 191-203.

[5] Bollerslev, T., 1986. Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics*. 31(3), 307–327.

[6] Bremnes, J.B., 2004. Probabilistic wind power forecasts using local quantile regression. *Wind Energy*. 7(1), 47–54.

[7] Bremnes, J.B., 2006. A comparison of a few statistical models for making quantile wind power forecasts. *Wind Energy*. 9(1-2), 3–11.

[8] Brier, G.W., 1950. Verification of forecasts in terms of probabilities. *Monthly Weather Review*. 78(1), 1-3.

[9] Brocker, J., Smith, LA., 2007. Scoring probabilistic forecasts: on the importance of being proper. *Weather and Forecasting*. 22(2), 382–388.

[10] Campbell, S.D., Diebold, F.X., 2005. Weather forecasting for weather derivatives. *Journal of the American Statistical Association*. 100(469), 6–16.

[11] Cannon, A.J., 2011. Quantile regression neural networks: Implementation in R and application to precipitation downscaling. *Computers and Geosciences*. 37, 1277-1284.

[12] Casati, B, Wilson, L.J., Stephenson, D.B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E.E., Brown, B.G., Mason, S., 2008. Forecast verification: current status and future directions. *Meteorological Applications* 15(1): 3–18.

[13] Chan, W.S., 1999. A comparison of some pattern identification methods for order determination of mixed ARMA models. *Statistics and Probability Letters*. 42(1), 69–79.

[14] Chang, Y., Park, J.Y., 2002. On the asymptotics of ADF tests for unit roots. *Econometric Reviews*. 21(4), 431–447.

[15] Chatfield, C., 1989. The analysis of time series, an introduction. Chapman and Hall, London.

[16] Chatfield, C., 1993. Calculating interval forecasts. *Journal of Business and Economics Statistics*. 11(2), 121-135.

[17] Chatfield, C., 2001. Prediction intervals for time-series forecasting, in: Armstrong, J.S., Principles of forecasting - A handbook for researchers and practitioners, Kluwer Academic, Amsterdam, pp. 475-494.

[18] Christoffersen, P.E., 1998. Evaluating prediction intervals. *Intern. Economic Review*. 39(4), 841-862.

[19] Cryer, J. D. and Chan, K. S. (2008). Time series analysis with applications in R. New York: Springer

[20] Ehrendorfer, M., 1997. Predicting the uncertainty of numerical weather forecasts: a review. Meteorologische Zeitschrift, Neue Folge, 6, 147–183.

[21] Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 50, 987–1007.

[22] Engle, R., 2001. GARCH 101: The use of ARCH/GARCH models in applied econometrics. *The Journal of Economic Perspectives*, 15(4), 157-168.

[23] Bach, F.R., Jordan, M.I., 2002. Kernel independent component analysis. *J. of Machine Learning Research*. 3, 1-48.

[24] Franses, P.H., Neele, J., Van Dijk, D., 2001. Modeling asymmetric volatility in weekly Dutch temperature data. *Environmental Modelling and Software*. 16(2), 131–137.

[25] Fuller, W.A., 1996. Introduction to statistical time series, 2nd ed., John Wiley & Sons, New York

[26] Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*. 102(477), 359–378.

[27] Good, I.J., 1952. Rational decisions. *Journal of the Royal Statistical Society*. 14, 107–114.

[28] Hagedorn, R., Smith, L.A., 2009. Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*. 16(2), 143–155.

[29] Hahn, G.J., Meeker, W.Q., 1991. Statistical intervals: A guide for practitioners, John Wiley, New York.

[30] Hansen, B.E., 1994. Autoregressive conditional density estimation. *International Economic Review*. 35(3), 705–730.

[31] Hansen, P.R., Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?. *Journal of applied econometrics*. 20(7), 873-889.

[32] Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models, Chapman and Hall, London.

[33] Hersbach, H., 2000. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570.

[34] Hosek, J., Musilek, P., Lozowski, E., Pytlak, P., 2011. Effect of time resolution of meteorological inputs on dynamic thermal rating calculations. *IET Generation, Transmission and Distribution*. 5(9), 941-947.

[35] Huth, R., Beck, C., Phillipp, A., Demuzere, M., Ustrnul, Z., Cahynova, M., Kysely, J., Tveito, O.E., 2008. Classifications of atmospheric circulation patterns: recent advances and applications. *Annals of the New York Academy of Sciences*. 1146, 105–152.

[36] Jain, A., Murty, M., Flynn, P., 1999. Data clustering: a review. *ACM Computing Surveys*. 31(3), 264–323.

[37] Jamali, A., Ghamati, M., Ahmadi, B., Nariman-zadeh, N., 2013. Probability of failure for uncertain control systems using neural networks and multi-objective uniform-diversity genetic algorithms (MUGA). *Engineering Applications of Artificial Intelligence*. 26(2), 714-723.

[38] Jooyoung, J., Taylor, J.W., 2012. Using conditional kernel density estimation for wind power density forecasting. *Journal of the American Statistical Association.* 107(497), 66-79.

[39] Jolliffe, I. T., and Stephenson, D. B. (eds.) (2003), Forecast verification: A practitioner's guide in atmospheric science, Chichester, U.K.: Wiley.

[40] Jørgensen, M., Sjøberg, D.I.K., 2003. An effort prediction interval approach based on the empirical distribution of previous estimation accuracy. *Information and Software Technology*. 45(3), 123-136.

[41] Kaufman, L., Rousueeuw, P.J., 1990. Finding groups in data: An introduction to cluster analysis, John Wiley & Sons, New York.

[42] Khaki, M., Musilek, P., Heckenbergerova, J., Koval, D., 2010. Electric power system cost/loss optimization using dynamic thermal rating and linear programming. In *Proceedings of the IEEE Electric Power and Energy Conference EPEC*.

[43] Koehler, A.B., 1990. An inappropriate prediction interval. *Int. J. Forecasting*. 6(4), 557-558.

[44] Koenker, R., 2005. Quantile regression. Cambridge University Press, Cambridge.

[45] Koenker, R., D'Orey, V., 1987. Computing regression quantiles. *Applied Statistics*. 36(3), 383–393.

[46] Landberg, L., 1999. Short-term prediction of the power production from wind farms. *Journal of Wind Engineering and Industrial Aerodynamics*. 80(1), 207-220.

[47] Lange, M., 2005. On the uncertainty of wind power predictions—Analysis of the forecast accuracy and statistical distribution of errors. *Journal of Solar Energy Engineering*. 127(2): 177–184.

[48] Lange, M., 2003. Analysis of the uncertainty of wind power predictions. PhD thesis, University of Oldenburg.

[49] Lange, M., Heinemann, D., 2003. Relating the uncertainty of short-term wind speed predictions to meteorological situations with methods from synoptic climatology. In

Proceedings of the European Wind Energy Conference, EWEC.

[50] Li, W.K., 2004. Diagnostic checks in time series. Chapman and Hall, London.

[51] Li, Y., Liu, Y., Zhu, J., 2007. Quantile regression in reproducing kernel hilbert spaces. *Journal of the American Statistical Association*. 102(477), 255–268.

[52] Lojowska, A., Kurowicka, D., Papaefthymiou, G., Van Der Sluis, L., 2010. Advantages of ARMA-GARCH wind speed time series modeling. In *Proceedings of the 11th IEEE International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*.

[53] Mazloumi, E., Rose, G., Currie, G., Moridpour, S., 2011. Prediction intervals to account for uncertainties in neural network predictions: Methodology and application in bus travel time prediction. *Engineering Applications of Artificial Intelligence*. 24(3), 534-542.

[54] Møller, J.K., Nielsen, H.A., Madsen, H., 2008. Time-adaptive quantile regression. *Computational Statistics & Data Analysis*. 52(3), 1292–1303.

[55] Mudelsee, M., 2010. Climate time series analysis: classical statistical and bootstrap methods. Springer, Dordrecht.

[56] Murphy, A.H., 1971. A note on the ranked probability score. *Applied Meteorology*. 10(1), 155-156.

[57] Murphy, A. H., 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*. 8(2), 281–293.

[58] Nelson, D.B., 1991. Conditional heteroscedasticity in asset returns: A new approach. *Econometrica*. 59, 347–370.

[59] Nielsen, H, Madsen, H, Nielsen, TS., 2006: Using quantile regression to extend an existing wind power forecasting system with probabilistic forecasts. *Wind Energy*. 9(1-2), 95–108.

[60] Nipen, T., Stull, R., 2011. Calibrating probabilistic forecasts from an NWP ensemble. *Tellus*. 63(5), 858–875.

[61] Orrell, D., Smith, L., Barkmeijer, J., Palmer, T., 2001. Model error in weather forecasting. *Nonlinear Processes in Geophysics*. 8(6), 357–371.

[62] Palmer, T.N., 2000. Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*. 63(2), 71-116.

[63] Pedrycz, W., 1990. Fuzzy sets in pattern recognition: Methodology and methods. *Pattern Recognition*. 23(1–2), 121-146.

[64] Pedrycz, W., 2005. Knowledge-based clustering: From data to information granules. Wiley, Hoboken, NJ.

[65] Pierolo, R., 2011. Information gain as a score for probabilistic forecasts. *Meteorological Application*. 18(1), 9-17.

[66] Pinson, P., 2006. Estimation of the uncertainty in wind power forecasting. PhD Dissertation, Ecole des Mines de Paris.

[67] Pinson, P, Juban, J, Kariniotakis, G., 2006. On the quality and value of probabilistic

forecasts of wind generation. In *Proceedings of the IEEE Conference on Probabilistic Methods Applied to Power Systems*.

[68] Pinson, P., Kariniotakis, G., 2010. Conditional prediction intervals of wind power generation. *IEEE Transactions on Power Systems*. 25(4), 1845-1856.

[69] Pinson, P., Nielsen, H.Aa., MZller, J.K., Madsen, H., Kariniotakis, G.N., 2007. Nonparametric probabilistic forecasts of wind power: required properties and evaluation. *Wind Energy*. 10(6), 497–516.

[70] Pytlak, P., Musilek, P., Lozowski, E., Arnold, D., 2010. Evolutionary optimization of an ice accretion forecasting system. *Monthly Weather Review*. 138(7), 2913-2929.

[71] Pytlak, P., Musilek, P., Lozowski, E., Toth, J., 2011. Modelling precipitation cooling of overhead conductors. *Electric Power Systems Research*. 81(12), 2147-2154.

[72] Richardson, D.S., 2000. Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*. 126(563), 649–667.

[73] Roulston, M.S., Smith, L.A., 2001. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*. 130(6), 1653–1660.

[74] Shumway, R.H., Stoffer, D. S., 2006. Time Series Analysis and Its Applications with R Examples, 2nd ed. Springer, New York.

[75] Silverman, B.W., 1986. Density estimation for statistics and data analysis. Chapman and Hall.

[76] Smola, A.J., 1998: Learning with kernels. Doctoral Dissertation, Technische Universitat Berlin.

[77] Takeuchi, I., Le, Q.V., Sears, T., Smola, A.J., 2006. Nonparametric quantile estimation. *Journal of Machine Learning Research*. 7, 1231-1264.

[78] Taylor, J.W., Buizza, R., 2004. Comparing temperature density forecasts from GARCH and atmospheric models. *Journal of Forecasting*. 23(5), 337–355.

[79] Taylor, J.W., McSharry, P.E., Buizza, R., 2009. Wind power density forecasting using ensemble predictions and time series models. *IEEE Transactions on Energy Conversion*. 24(3), 775-782.

[80] The Weather Research and Forecasting (WRF) Model: http://www.wrf-model.org/index.php, accessed on December 2012.

[81] Toth, Z., 2003. Probability and ensemble forecasts, in Jolliffe, I.T., Stephenson, D.B., Forecast verification: A practitioner's guide in atmospheric science, Wiley & Sons, New York, pp. 137–164.

[82] Tsay, R.S., 1984. Regression models with time series errors. *Journal of the American Statistical Association*. 79(385), 118–24.

[83] Vapnik, V., 1998. Statistical learning theory. Wiley, New York.

[84] Vejmelka, M., Musilek, P., Palus, M., Pelikan, E., 2009. K-means clustering for problems with periodic attributes. *International Journal of Pattern Recognition and Artificial Intelligence*. 23(4), 721-743.

[85] Visser, H., Molenaar, J., 1995. Trend estimation and regression analysis in climatological time series: An application of structural time series models and the kalman filter. *Journal of Climate*, 8(5), 969–979.

[86] Wilks, D.S., 2006. Statistical methods in the atmospheric sciences. Academic Press, New York.

[87] Winkler, R.L., 1972: A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*. 67(337), 187–191.

[88] Wonnacott, T.H., Wonnacott, R.J., 1990. Introductory statistics. Wiley, New York.

[89] Xu, R., Wunsch II, D., 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. 16(3), 645–678.

[90] Yu, K., Zudi, L., 2003. Quantile regression: Applications and current research areas. *Journal of the Royal Statistical Society*. 52(3), 331-350.

[91] Yu, K., Jones, M.C., 1998. Local linear quantile regression. *Journal of the American Statistical Association*. 93(441), 228–238.