Using a Bayesian regression approach on dual-model wind storm simulations to improve wind speed prediction

Jaemo Yang, Marina Astitha, Emmanouil N. Anagnostou

Department of Civil and Environmental Engineering

University of Connecticut, Storrs, CT, USA

Brian M. Hartman

Department of Statistics

Brigham Young University, Provo, UT, USA

Submitted to: Journal of Applied Meteorology and Climatology

Submission date: October 28, 2016

Corresponding Author: Marina Astitha, Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT 06269-3037. Phone: (860) 486-3941. Email: astitha@engr.uconn.edu

Abstract

Weather prediction accuracy is very important given the devastating effects of extreme weather events in recent years. Numerical weather prediction (NWP) systems are used to build strategies to prevent catastrophic losses of human lives and the environment and have evolved with the use of multi-model or single-model ensembles and data assimilation techniques in an attempt to improve the forecast skill. However, these techniques require increased computational power (thousands of CPUs) due to the number of model simulations and ingestion of observational data from a wide variety of sources.

In this study, the combination of predictions from two state-of-the-science atmospheric models (WRF and RAMS/ICLAMS) using Bayesian and simple linear regression techniques is examined, and the improvement in wind speed prediction for the Northeast United States (NE U.S.) using regression techniques is demonstrated. Retrospective simulations of seventeen storms that affected NE U.S. during the period 2004-2013 are performed and utilized. Optimal variances are estimated for the thirteen training storms by minimizing the root mean square error and are applied to four out-of-sample storms (Hurricane Irene (2011), Hurricane Sandy (2012), November 2012 winter storm and February 2013 blizzard). The results show a 20-30% improvement in the systematic and random error of 10-m wind speed over all stations and storms, using various storm combinations for the training dataset. This study indicates that 10 to 13 storms in the training dataset are sufficient to reduce the errors in the prediction and a selection based on occurrence (chronological sequence) is also considered efficient.

1. Introduction

Weather forecasting, applied to global and regional scales, has evolved with the use of multi-model or single-model ensembles (Doblas-Reyes et al. 2005; Palmer et al. 2008; Weigel et al. 2009; Kirtman et al. 2014), data assimilation techniques (Barker et al. 2012; Wang et al. 2013; Ancell et al. 2015) and high-resolution grid spacing (Roberts 2003; Speer et al. 2003; Steppeler et al. 2003; Gego et al. 2005; Schwartz et al. 2009) in an attempt to improve the forecast skill. Despite the noted improvements, inaccuracies caused by random and systematic errors are a continuous topic for research (Krishnamurti et al. 2004; Mass et al. 2008; Ancell et al. 2011; 2012; Delle Monache et al. 2011). The ability of numerical weather prediction (NWP) models to accurately describe atmospheric conditions under various dynamic states is influenced by errors caused from the implemented physical parameterizations, initial state, boundary conditions and data availability. Atmospheric complexity and inability to handle sub-grid scale phenomena also cause errors in the predicted meteorological variables (Libonati et al. 2008; Louka et al. 2008; Idowu and deW Rautenbach 2009). Restrictions in the resolution cause the imperfect representation of the actual surface properties (e.g., topography, vegetation, soil types and moisture, and sea surface temperature) which can result in significant model error along the sharp gradients. In addition, inaccurate prediction of land surface interactions can be disadvantageous to the NWP (Koster and Suarez 2001; Drusch and Viterbo 2007; Papadopoulos et al. 2008; Serpetzoglou et al. 2010), and approximation of the planetary boundary layer representation can be an error source for the prediction of surface variables (Arakawa 2004; Pleim 2007; Hu et al. 2010; Nielsen-Gammon et al. 2010; Frediani et al. 2016).

Using high spatial resolution and/or data assimilation does not always assure high forecast accuracy, because of the important role of the input fields, initial conditions and inherent model uncertainties that influence the prediction. Statistical post-processing approaches play a useful role to address this issue and contribute to the reduction of prediction errors. Various techniques on statistical post-processing for error/bias correction have been suggested in the literature. These techniques are based on: (1) running mean bias removal (Stensrud and Yussouf 2003; 2005; Eckel and Mass 2005; Hacker and Rife 2007; Wilczak et al. 2006) (2) Kalman filter (KF) post-processing (Libonati et al. 2008; Müller 2011; Homleid 1995; Roeger et al. 2003; McCollor and Stull 2008; Rincon et al. 2010; Delle Monache et al. 2006; 2008; 2011; Djalaova et al. 2010; Kang et al. 2010) (3) Model Output Statistics (MOS) (Glahn and Lowry 1972; Carter et al. 1989; Jacks et al. 1990; Mao et al. 1999; Wilson and Vallée 2002; 2003; Hart et al. 2004; Wilks and Hamill 2007; Glahn et al. 2009).

Combining statistical post processing techniques with NWP ensemble simulations is of particular interest due to the ability to characterize model uncertainty and improve the predicted variables (e.g., wind speed, temperature, humidity, etc.). Even though there is no consensus on the adequate amount of ensemble members as well as the best way to combine them (Weigel et al. 2010), computational cost of ensemble simulations can be a deterrent factor. The motivation for this work has its basis on the use of a computationally efficient scheme that uses only two NWP models and statistical post-processing techniques over a set of meteorological storms that have common characteristics. One typical method for optimum weighting of the ensemble members is Bayesian model averaging (BMA) that estimates each member's contributing weight (Raftery et al. 2005; Wilson et al. 2007; Fraley et al. 2010; Erickson et al. 2012; Sloughter et al. 2007; 2010). In the BMA approach, the calibrated weights reflect the forecast skill of each ensemble member over a training period (Fraley et al. 2010). Fraley et al. (2010) implemented BMA with 86 members, which represent a relatively large number of ensemble members, to show how the BMA can be adapted to handle exchangeable ensemble members. Erickson et al. (2012) ran BMA for specific weather storms including fire weather that caused poor air quality. This test was conducted using sequential (the most recent days) and conditional training periods (the most recent similar days), and showed that the correction of conditional training period was better than the sequential training.

The similarity or difference between training and out-of-sample conditions can affect the results of statistical post-processing methods that accompany training algorithms. Although statistical post-processing methods correct the errors over general cases, specific storms may not be improved when the training period does not consider similar patterns with the target storms. In other words, if the training dataset reflects the characteristics of the target storm, the modeled field may be improved more efficiently. Especially for high wind speed weather storms, distinguishing the mean atmospheric conditions and using a training scheme with a dataset fitted to similar weather conditions can be a critical factor for the success of the error correction. Our reference to extreme storms includes tropical storms, heavy precipitation associated with floods, blizzards with strong sustained winds, and seasonal thunderstorms.

The main objective of this study is to improve surface wind speed prediction under extreme weather conditions, as it is strongly correlated with negative effects in civil infrastructure, power grid and the environment. To this end, the combination of wind speed predictions from two atmospheric models using a Bayesian Linear Regression (BLR) approach is explored, and the potential to improve wind speed prediction against single model simulations and Simple Linear Regression (SLR) techniques is demonstrated. The combination of two atmospheric modeling systems with simple bias correction techniques serves two purposes: minimizes computational cost since only two model members are being employed and determines the value added by Bayesian regression in a deterministic framework. An additional goal of this work is to assess the efficient length for the training period in chronological and non-chronological sequences, which will be important in the operational application of the described methodology. The work presented here will support the operational prediction of power outages in NE U.S. that are strongly influenced by wind severity (Wanik et al. 2015; He et al. 2016). Currently, the power outage modeling system is operating with meteorological inputs from the WRF model (Wanik et al. 2015). Section 2 describes the model configuration and data used, section 3 presents the methodology for SLR and BLR and section 4 includes discussion of the results. Conclusions and future work are summarized in section 5.

2. Models and data

a. Atmospheric modeling systems

Two mesoscale meteorological modeling systems are implemented to simulate the selected storms. The Weather Research and Forecasting model (WRF-ARW version 3.4.1; referred to as WRF) (Skamarock et al. 2008) and the Regional Atmospheric Modeling System/Integrated Community Limited Area Modeling System (RAMS/ICLAMS; referred to as ICLAMS) (Cotton et al. 2003; Solomos et al. 2011; Kushta et al. 2014). ICLAMS is an integrated air quality and chemical weather modeling system based on RAMSv6 (Pielke et al. 1992; Cotton et al. 2003) that directly couples meteorological fields with air quality components, and includes gaseous, aqueous, aerosol phase chemistry and partitioning of cloud condensation nuclei (CCN), giant cloud condensation nuclei (GCCN), and ice nuclei (IN) as predictive quantities (atmospheric chemistry and feedback processes are not included in the ICLAMS simulations for this work).

Both models have three nested domains covering the Northeast U.S. with horizontal grid spacing of 18 km (outer domain), 6 km (inner-intermediate domain) and 2 km. The third gridded domain is the focus area in this work (Fig. 1a, b). To initialize the two models, the National Centers for Environmental Prediction (NCEP) Global Forecast System ($1^{\circ} \times 1^{\circ}$, 6-hourly intervals) analyses (NCEP/NOAA, 2007) and the Final Analysis ($1^{\circ} \times 1^{\circ}$, 6-hourly intervals) data (NCEP/NOAA, 2000) are used for WRF and ICLAMS respectively. Configuration details for both WRF and ICLAMS are summarized in Table 1.

The storms that comprise the training and validation datasets are selected after a k-mean clustering analysis of sea level pressure, 2-m temperature and 10-m wind speed from the European Centre for Medium-Range Weather Forecasts (ECMWF) Interim Re-Analysis (ERA-Interim: Simmons et al. 2007) for 80 storms that affected the power network in NE U.S. (from 20 outages to > 15,000 outages) and span the period 2004-2013 (Maria Frediani, personal communication, 2015). A subset of seventeen storms is selected, which belong to two clusters representing winter and late summer/fall season storms with strong winds and intense pressure gradients. The selected storms include three major storms for NE U.S.: Hurricane Irene (2011), Hurricane Sandy (2012) and the 8-9 February (2013) blizzard. General information about the storms is included in Table 2.

b. Observations

The Automated Surface Observing System (ASOS) observation datasets at the National Centers for Environmental Prediction (NCEP) are used for model evaluation and also for the implementation of error optimization (SLR and BLR). The ASOS generally provides minute-by-minute observations and generate the Meteorological Terminal Aviation Routine Weather Report (METAR) and Aviation Selected Special Weather (SPECI) report. ASOS is installed at more than 900 airports across the United States, and the data from 80 stations over the Northeast U.S. are used in this study (Fig. 1c). The wind speed at observational locations are matched with the modeled wind speed using bilinear interpolation (nested grid at 2 km $\times$ 2 km grid spacing).

3. Methodology

Two statistical post-processing methods, the Simple Linear Regression (SLR) and Bayesian Linear Regression (BLR) are applied for error correction of the modeled wind speed. Thirteen storms are used for the training dataset, and four storms for the out-of-sample application (validation). The first application uses a chronological sequence to select the storms for the training dataset. The second application uses all possible combinations of the thirteen storms to compose the training dataset, regardless of the date of occurrence. The two regression methods are described in sections 3a and b; section 3c presents details on the training scheme and section 3d includes information about data processing and statistical metrics.

a. Simple Linear Regression (SLR)

The SLR model consists of the mean and variance function (Weisberg 2005), defined as follows:

$$E(Y|X=x)=\beta_0+\beta_1 x \tag{1}$$

Intercept β_0 is the value of $E(Y|X=x)$ when x equals zero, and the slope β_1 is the rate of change in $E(Y|X=x)$ for a unit change in X, respectively. The unknown parameters β_0 and β_1 are estimated from the modeled-observed wind speed pairs given the independent and dependent variable vector, X and Y.

In this study, the SLR model with a single predictor (wind speed at 10m) is developed for WRF and ICLAMS separately. To evaluate the estimators, storms are selected to arrange the training datasets first, and each estimator β_0 and β_1 is calculated from the training storms by the ordinary least squares (OLS) method as follows:

$$\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) / \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

where $\bar{x}$, $\bar{y}$ are the averages of modeled values x and observed values y in the training datasets. Since a linear relationship between observed and modeled wind speed exists, this relationship points towards the possibility of model prediction correction (Sweeney et al. 2013). The SLR method is developed for each station because the linear relationship is spatially variable (station to station) and the spatial error heterogeneity must be preserved in the results. Therefore, the SLR analysis is implemented for each station (total of 80 stations) by the OLS method, and overall the final SLR model for WRF and ICLAMS is given as:

$$\hat{Y}_{station\ m} = \hat{\beta}_{0,\ \ station\ m} + \hat{\beta}_{1,\ \ station\ m} X_{station\ m} \tag{4}$$

where $\hat{\beta}_{0,\ \ station\ m}$, $\hat{\beta}_{1,\ \ station\ m}$ are the estimators of station m evaluated from the training dataset, and $X_{station\ m}$ is the predictor of the station m from the WRF or ICLAMS out-of-sample storms. $\hat{Y}_{station\ m}$ is the final product of the SLR model for station m.

b. Bayesian Linear Regression (BLR)

BLR is implemented as a new approach to improve WRF and ICLAMS 10-m wind speed fields. Bayesian statistics are based on Bayes' theorem that deals with uncertainty of unknown parameters, and the parameters are basically inferred to probabilistic forms from the observed data under the Bayesian framework (Chu and Zhao 2011). The Bayesian formula to infer the unknown parameter vector θ is thus governed by:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \tag{5}$$

where $p(\theta|y)$ is the posterior probability density function (PDF) of θ given the observed data information y, $p(y|\theta)$ is the likelihood function, $p(\theta)$ is the prior PDF for the unknown parameter vector of θ, and $p(y)$ is the PDF of the observation vector y. Considering the continuous case for the Bayes' theorem, Eq. (5) is formulated as follows:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)\,p(\theta)d\theta} \tag{6}$$

In this study, two controlled variables produced from WRF and ICLAMS are used in the BLR, so the normal linear model is regarded as a normal multiple regression form defined by two predictor variables. In vector form, the normal multiple regression equation is defined by:

$$y = \beta X + \varepsilon \tag{7}$$

where: y is a $n \times 1$ vector of observations; X is a $n \times p$ matrix of independent variables incorporating the unit matrix of the first column for the intercepts β_0; β is a $p \times 1$ vector of regression coefficients $(\beta_0, \ \beta_1, \ \beta_2)$; and the error term is $\varepsilon \sim N(0, I\sigma^2)$ with the unknown dispersion parameter σ^2. After considering the elements of the parameter vector $\theta = [\beta_0, \beta_1, \beta_2, \sigma^2]$, Eq. (6) becomes:

$$p(\beta_0, \beta_1, \beta_2, \sigma^2|y) = \frac{p(y|\beta_0, \beta_1, \beta_2, \sigma^2)p(\beta_0, \beta_1, \beta_2, \sigma^2)}{\int\int\int\int p(y|\beta_0, \beta_1, \beta_2, \sigma^2)p(\beta_0, \beta_1, \beta_2, \sigma^2)d\beta_0 d\beta_1 d\beta_2 d\ \sigma^2} \tag{8}$$

It is assumed that the posterior probability distribution is in the same family as the prior probability distribution and prior information of σ^2 can be inferred. The posterior mean of β can be calculated by Eq. (9) (Cattin et al. 1983; Zellner 1996; O'Hagan 2004; Sorensen and Gianola 2002; Walter et al. 2007; Walter and Augustin 2010):

$$\bar{\beta} = \begin{bmatrix} \bar{\beta}_0 \\ \bar{\beta}_1 \\ \bar{\beta}_2 \end{bmatrix} = \left(X^T X + V_\beta^{-1}\right)^{-1} (X^T y + V_\beta^{-1} \mu_\beta) \tag{9}$$

$$y = \begin{bmatrix} Obs_1 \\ Obs_2 \\ \vdots \\ Obs_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{1_{WRF}} & x_{1_{ICLAMS}} \\ 1 & x_{2_{WRF}} & x_{2_{ICLAMS}} \\ \vdots & \vdots & \vdots \\ 1 & x_{n_{WRF}} & x_{n_{ICLAMS}} \end{bmatrix} \qquad V_\beta = \begin{bmatrix} \sigma_0^2 & 0 & 0 \\ 0 & \sigma_1^2 & 0 \\ 0 & 0 & \sigma_2^2 \end{bmatrix} \quad \mu_\beta = \begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \end{bmatrix}$$

where: y is the matrix of 10-m observed wind speed for *n* time steps; X is the matrix of WRF and ICLAMS wind speed for *n* time steps; V_β is a diagonal matrix including three prior variances which correspond to each element of $\bar{\beta}$; and μ_β is a prior mean matrix. It is assumed a *priori* that the best model will be a simple unbiased average of the two simulations, implying a mean vector of $\mu_\beta = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}$. Certainty about that assumption is defined by the size of the prior variances (σ_i^2). To rely more on the data to inform the final model, the prior variances are made much larger. In an extreme case, as the prior variances go to infinity, Bayesian posterior estimates will match the OLS estimates. As the prior variances get smaller, the results will shrink toward the *a priori* assumptions. Shrinkage of this type will allow the model to be more robust in the presence of outliers and other strange and influential data. In order to develop BLR using Eq. (9), optimal prior variances are searched with the matrix representing the prior mean (μ_β). The final BLR model based on WRF and ICLAMS is formulated as follows:

$$\hat{Y}_{station\ m} = \bar{\beta}_{0,\ \ station\ m} + \bar{\beta}_{1,\ \ station\ m} X_{WRF,\ \ station\ m} + \bar{\beta}_{2,\ \ station\ m} X_{ICLAMS,\ \ station\ m} \tag{10}$$

where $\bar{\beta}_{0,\ \ station\ m}$ is the intercept of the BLR equation, $\bar{\beta}_{1,\ \ station\ m}$ and $\bar{\beta}_{2,\ \ station\ m}$ denote the regression coefficients for the two predictor variables $X_{WRF,\ \ station\ m}$ and $X_{RAMS/ICLAMS,\ \ station\ m}$ for station m. $\hat{Y}_{station\ m}$ is the adjusted 10-m wind speed field for station m using the BLR method.

c. Training scheme

The regression coefficients for SLR and BLR can be estimated using a variety of training datasets. To investigate the sensitivity of the results to training period length, a variation in the number of storms as well as a change in the chronological order of the training dataset are examined. For instance, in the case of using one storm for the training dataset, SLR is implemented for each station, and then the number of storms is gradually increased to make different combinations. Among the seventeen storms, thirteen storms from 2004 to 2011 are selected as training storms and the other four storms are used for out-of-sample applications/validations, respectively. The four storms represent significant storms over the Northeast U.S. during 2011 to 2013: Hurricane Irene (2011), Hurricane Sandy (2012), November 2012 storm (affected by Hurricane Sandy) and February 2013 blizzard (maximum 1-hour wind speed from 80 inland stations: 21, 25, 22, 24 m/s for Irene, Sandy, November 2012 storm and February 2013 blizzard).

In the first approach, an increasing number of storms in chronological order are employed. The second approach consists of training datasets composed of all possible combinations of the thirteen storms to analyze the behavior of R^2, RMSE, BIAS and CRMSE in conformity with changes in combinations. Each quantity of combination using different number of storms can be calculated by $n!/(r!\,(n-r)!)$ which represents the number of r-combinations from a given set of n elements (r is an integer; and $1 \leq r \leq 13,\ n = 13$). Specifically, in the case of using a single storm, thirteen individual storms constitute the training dataset and, in the case of using two storms, seventy eight training datasets are required (Table 3). The experiments for all possible combinations (with a total of 8191 training datasets, Table 3) are implemented for each observation station for SLR and BLR.

The BLR approach for the first application has three phases (Fig. 2): **(1)** Random selection of 10,000 prior variance sets (V_{β}, Eq. 9) within the interval $[10^{-10}, 1]$. Each variance set (out of the 10,000) is used to estimate $\bar{\beta}$ (Eq. 9) for each station using all training storms. The estimated $\bar{\beta}$ is applied to individual storms of the training dataset to compute the global RMSE for each storm and each station (Phase 1 in Fig.2). **(2)** The RMSE that corresponds to each variance set is summed over all k storms (Phase 2 in Fig.2). **(3)** The optimal prior variances that corresponded to the minimum summation of the RMSE from phase 2, are used for the calculation of the final $\bar{\beta}$ for each station which is then applied to out-of-sample storms (Phase 3 in Fig. 2).

The BLR procedure demands a relatively longer computation time than SLR since it incorporates a phase to sample 10,000 prior variance sets. To reduce the computational time for BLR experiments related to all possible storm combinations, it is necessary to use fewer random samples of variance sets, instead of the previous 10,000. The variation of the RMSE from all cases, employing one up to thirteen storms for the training dataset, is analyzed to determine the reasonable number of prior variance samples that would reduce the computational cost and succeed in minimizing the RMSE similar to the 10,000 variance sets.

RMSE variability with increasing sample size is quantified by calculating the normalized difference of RMSE (NDiff) from the final minimized RMSE of all 10,000 samples. The normalized difference of RMSE is calculated as follows:

$$NDiff = \frac{\left[\min\limits_{i \in j}\left(\sum_{n=1}^{k} RMSE_{storm\ n,\quad i}\right), 1 \le j \le 10{,}000\right] - \min\limits_{i \in 10000}\left(\sum_{n=1}^{k} RMSE_{storm\ n,\quad i}\right)}{\min\limits_{i \in 10{,}000}\left(\sum_{n=1}^{k} RMSE_{storm\ n,\quad i}\right)} \times 100\% \tag{11}$$

where: i is the number of variance sets in the range of 1 and 10000; j is the number of variance sets to be used for reduction of the computational cost in the range of 1 and 10000; k is the number of storms.

For example, if 2 storms and 20 variance sets are used to calculate NDiff:

$$NDiff = \frac{\left[\min\limits_{i \in 20}\left(\sum_{n=1}^{2} RMSE_{storm\ n,\quad i}\right)\right] - \min\limits_{i \in 10000}\left(\sum_{n=1}^{2} RMSE_{storm\ n,\quad i}\right)}{\min\limits_{i \in 10000}\left(\sum_{n=1}^{2} RMSE_{storm\ n,\quad i}\right)} \times 100\%$$

All cases commonly display that the normalized difference of RMSE values are decreased near the 20 variance sets (Fig. 3). Thus, 20 samples are identified as a proper sample size for prior variance sets and are used to implement BLR for the all-storm combinations.

d. Data processing and statistical metrics

The first 6 hours are regarded as the model spin-up time and are discarded from the analysis. The missing and zero values for 10-m wind speed observations are not included in the modeled-observed pairs. Five statistical metrics that offer complementary views on the model and regression performances are used. To evaluate the impact of regression techniques on the 10-m modeled wind speed, the metrics are calculated for raw WRF, raw ICLAMS, WRF_{SLR}, $ICLAMS_{SLR}$ and BLR. The five statistical metrics used in this study are: coefficient of determination (R^2), root mean square error (RMSE), mean bias (BIAS), centered root mean square error (CRMSE; Murphy 1988; Taylor 2001; Delle Monache et al. 2011), and skill score (SS) which are determined as follows:

$$R^2 = \left[\frac{N\sum_N (X\cdot Y) - (\sum_N X)(\sum_N Y)}{\sqrt{[N\sum_N X^2 - (\sum_N X)^2][N\sum_N Y^2 - (\sum_N Y)^2]}}\right]^2 \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_N (X-Y)^2} \tag{13}$$

$$BIAS = \frac{1}{N}\sum_N (X-Y) \tag{14}$$

$$CRMSE = \sqrt{\frac{1}{N}\sum_N [(X-\bar{X})-(Y-\bar{Y})]^2} \tag{15}$$

where: the modeled value is represented by X, the observed wind speed by Y, N is the total number of data points, and $\bar{X}$ and $\bar{Y}$ are the modeled and observed wind speed averages over the N values used in the calculations. RMSE is used to evaluate model performances and the crucial objective function aiming to mitigate errors for the BLR approach. CRMSE is a measure of the random component of RMSE, while the systematic component is represented by the BIAS.

To measure the relative improvement of the regression techniques, the Skill Score (SS) with regards to RMSE and R^2 (e.g., Wilks 1995; Libonati et al. 2008; Idowu and deW Rautenbach 2009; Delle Monache et al. 2011) is calculated. An example of the SS calculation is shown in Eq. (16) and (17):

$$SS_{RMSE} = \frac{RMSE_{raw} - RMSE_{SLR/BLR}}{RMSE_{raw}} \times 100\% \tag{16}$$

$$SS_{R^2} = \frac{{R^2}_{SLR/BLR} - {R^2}_{raw}}{{R^2}_{raw}} \times 100\% \tag{17}$$

Eq. (16) and Eq. (17) estimate the relative improvement of the SLR and BLR approaches versus raw-WRF and raw-ICLAMS predictions. Positive values of SS_{RMSE} and SS_{R^2} indicate that the suggested regression method improves the raw model outputs.

4. Results and discussion

a. Chronologically-ordered storm combinations for the training dataset

The variation of R^2, RMSE, BIAS and CRMSE (Figs. 4 and 5) for each out-of-sample storm shows an increase (R^2)/decrease (RMSE, BIAS and CRMSE) when the number of storms of the training dataset increases. All three models (WRF_{SLR} (triangles), $ICLAMS_{SLR}$ (squares) and BLR (circles)) exhibit poor performance indicated by low R^2 and high RMSE, BIAS and CRMSE when employing one storm for training, which denotes that one historical storm is not sufficient to improve wind speed predictions of future storms. The statistical metrics progressively improve with increasing number of storms in the training dataset, and the trend reaches a plateau after eight to ten storms to an almost constant value. This is indicative of the number of storms that will be efficient and effective for correcting the modeled wind speed error. BLR is consistently performing better across all storm cases with the only exception of the November 2012 storm, where BLR and $ICLAMS_{SLR}$ share comparable performances. 95% bootstrapped confidence intervals are included for all statistical metrics and out-of-sample storms (shown in Figs. 4,5). Non-overlapping bootstrapped intervals show that results are significantly different when looking at the RMSE for Irene and Sandy, for the maximum number of storms in the training dataset. For the latest two out-of-sample storms,

$ICLAMS_{SLR}$ and BLR are not significantly different in terms of the RMSE.

The mean bias is almost entirely removed for most out-of-sample storms with all models being successful (Fig. 5). The mean bias of raw model outputs for the four storms is in the range of -1.0 m/s and 0.5 m/s. At least 5 storms are required for the success of bias removal by SLR and BLR (Fig. 5). Hurricane Sandy wind speed exhibits a positive bias even with the inclusion of 13 storms in the training dataset and BLR has higher bias than WRF or ICLAMS. This is attributed to the fact that model predictions of Sandy exhibit distinctly different error characteristics than the other storms in the database. To explore this behavior further, Hurricane Irene was included in the training dataset (14 storms instead of 13) as the storm closest in character to Hurricane Sandy. Sandy, November 2012 storm and February 2013 blizzard are kept as the out-of-sample storms and the results did not show significant differences (not shown here). The average RMSE for Sandy changed only by 0.01 m/s and the spatial distribution was not significantly affected. The case of Sandy shows that the mean bias can be reduced when the available number of training storms increases. Bias removal is consistent with the systematic error removal that is an expected outcome of regression techniques. The random component of RMSE, denoted by centered RMSE (CRMSE), has a decreasing trend for all storms as the number of training storms increase (Fig. 5). For both RMSE and CRMSE, BLR results are more successful than SLR, with the exception of November 2012 storm (as previously noted).

So far, results from the regression techniques are discussed without mentioning the "raw" atmospheric model performance. The correlation increases to 0.6-0.8 and the RMSE decreases to 1.7-2.0 m/s, for the different out-of-sample storms. Distribution of weights [beta coefficients, Eq. (9)] given at NWP models in the BLR approach (including 13 storms in the training dataset) shows a slight "preference" towards the ICLAMS model (Fig. 6). These beta values give the optimal RMSE in the training and are subsequently applied to all out-of-sample storms but vary for each station. To put things in perspective, statistical metrics employing the raw model outputs are presented using the skill score (SS).

The skill score (SS; the relative improvement in percent for a given metric), grouped by storm and raw model output, indicates improvement by BLR compared to SLR (Table 4), marked by increased SS values for BLR versus raw model outputs for all storms. This indicates that the BLR approach has been successful in improving the RMSE and R^2 statistical metrics for wind speed compared to raw model outputs. Additionally, the RMSE and R^2 are analyzed for each model normalized by BLR, in order to identify how BLR performs over the other methods. Normalized RMSE values greater than one indicate that BLR performs better (all normalized RMSE values are greater than one with the exception of November 2012 storm and $ICLAMS_{SLR}$; Table 5). Inversely, normalized R^2 smaller than one, demonstrates that BLR outperforms the other model results in the out-of-sample storms. These normalized RMSE and R^2 values listed in Table 5, show the same patterns in terms of the BLR performance. Normalized metrics indicate that BLR improves the wind speed statistical metrics for Irene, Sandy and the February 2013 storm. $ICLAMS_{SLR}$ performs as close, if not better, than BLR for the November 2012 storm. The results from the normalized metrics are consistent with the conclusions from the confidence intervals discussed previously.

Finally, the spatial distribution of RMSE was analyzed, with thirteen storms as training dataset for all out-of-sample storms (Fig. 7). In each plot, colored circles represent the RMSE value calculated using obervations at each station location. All suggested regression methods in this study sucessfully reduce the RMSE of the raw WRF and raw ICLAMS for almost all stations and storms. The RMSE in raw model outputs ranges between 1.6 and 3.5 m/s, whereas using regression techniques, a large number of stations show decreased values within a range of 1.0 to 2.5 m/s (more abundance of lower range RMSE values). Overall, BLR is shown to be an effective method to reduce RMSE for the stations of our case study, with highest reductions in the range of 17%-32% when compared to raw model outputs. More details on the timeseries and RMSE values for individual stations are provided in Table S1 and Figs. S1-S4 pf the supplement. In addition, spatial distrbution of BIAS and CRMSE is provided in Figs. S5 and S6 of the supplement, for a more detailed view of the BLR efficiency at the station level.

b. All-storm combinations for the training dataset

In this section, the training dataset comprises of all possible storm combinations while increasing the number of storms. The intention of this test is to define the sensitivity of BLR and SLR results to a random combination of storm sequences and denote the confidence that can be placed in the BLR method if a convergence in the results is achieved. The results showing 25th, 50th, 75th percentiles (horizontal bars), minimum and maximum (error bars) (Fig. 8 and 9) are similar to the chronologically ordered selection of training storms in the sense that the bias is almost entirely removed in most cases, and RMSE and CRMSE are decreased with the addition of storms in the training dataset. The variability of all metrics is clearly reduced by adding more storms in the training dataset (box whiskers plots in Fig. 8 and 9). Even at the combination of six or seven storms (largest number of combinations=1716, Table 3), the distribution is narrow. BLR starts with a relatively narrow width distribution compared to the other models, having higher R^2 values and lower RMSE and CRMSE. For example, using a single training storm in the case of Irene (Fig. 8), statistical metrics for BLR exhibit the following ranges: R^2=[0.73, 0.78], RMSE=[1.91, 2.17] (m/s) and CRMSE=[1.91, 2.10] (m/s). These ranges are narrower than those of WRF_{SLR} (R^2=[0.59, 0.74], RMSE=[2.07, 2.59] (m/s) and CRMSE=[2.07, 2.52] (m/s)) and ICLAMS$_{SLR}$ (R^2=[0.65, 0.72], RMSE= [2.19, 2.92] (m/s) and CRMSE=[2.16, 2.79] (m/s)). In addition, the median values corresponding to BLR (R^2: 0.76, RMSE: 2.05 m/s and CRMSE: 1.98 m/s) indicate statistically significant improvements when compared to the WRF$_{SLR}$ (R^2: 0.68, RMSE: 2.28 m/s and CRMSE: 2.26 m/s) and ICLAMS$_{SLR}$ (R^2: 0.68, RMSE: 2.32 m/s and CRMSE: 2.31 m/s).

When the lowest possible RMSE is selected (minimum RMSE in the 12 storm combinations, Fig. 8 and 9) and used for the calculation of BLR weighting factors for the out-of-sample application, there is no significant change in the average RMSE over all stations, neither in the spatial distribution of RMSE (not shown). The results from combining all possible storm sequences denotes a convergence in the wind prediction improvements by both chronological and all-combinations approach, giving confidence on the performance of the proposed BLR technique.

6. Conclusions

In this study, a simple linear regression (SLR) and a Bayesian linear regression (BLR) are introduced as post prediction error correction techniques, to improve modeled 10-m wind speed of storms that exhibit high wind speed occurrences. Both simple and Bayesian linear regressions rely on the training dataset and the appropriate selection of storms with similar weather characteristics. A selection of seventeen storms in total are used to study the efficiency of the two methods in reducing the wind speed systematic and random errors for station locations in Northeast United States (NE U.S.). Thirteen storms constitute the training dataset and four high impact storms (two hurricanes, one blizzard and one nor'easter) are used for the out-of-sample applications.

Both SLR and BLR reduce systematic and random errors for most out-of-sample storms. The statistical metrics and spatial distribution of root mean square error (RMSE) indicate that BLR is more successful in the surface wind speed error correction as it takes into account wind predictions from two atmospheric modeling systems. Such result is promising because the two-model application reduces the computational cost associated with multi-model or single-model ensemble forecasts without compromising the accuracy of the wind speed error reduction.

The selection of storms in the training dataset does not depend on the chronological sequence of storm occurrence but mostly on their abundance. The randomized experiment shows a good convergence of wind speed forecast improvements for all possible storm combinations, increasing the confidence in the proposed BLR technique. A suggestion that applies to the specific type of weather storms included in this work, is that ten to thirteen storms in the training dataset are sufficient to reduce the errors in the prediction by 20-30% for all stations compared to raw model outputs (Table 4) and up to 60% for individual stations (Fig. S7 in the supplement). A selection based on occurrence (chronological sequence) is also considered sufficient. This conclusion allows for planning of real-time operational wind speed error correction using the BLR technique.

Overall, this study has demonstrated that the application of two regression methods can improve the surface wind speed prediction from single and dual-model simulations. The dual-model combination of the BLR approach is more skillful and merits further investigation. Future extensions of this work include distribution of optimized BLR coefficients to each grid point of the model domain to improve the modeled wind speed for all locations. Furthermore, beta-testing will be expanded to an operational set-up in the NE U.S. region, where the real-time wind speed prediction of a storm using the two-modeling system will be corrected based on historic storms included in the training dataset. This will be accomplished by running operationally both numerical weather prediction (NWP) models (WRF and RAMS/ICLAMS) daily with a 5-day forecast window (WRF is currently operational). The current practice of identifying a potential future storm by consulting the in-house NWP as well as other operational forecasts (e.g., National Weather Service (NWS), National Centers for Environmental Prediction (NCEP)), will be implemented, in which event, BLR will be applied to provide optimal dual-model wind speed predictions.

Acknowledgements

The work was supported by Eversource Energy through a research grant awarded by the Eversource Energy Center at the University of Connecticut. WRF is developed and maintained by the National Center for Atmospheric Research, funded by NSF.

References

Ancell, B. C., C. F. Mass, and G. J. Hakim, 2011: Evaluation of surface analyses and forecasts with a multiscale ensemble Kalman Filter in regions of complex terrain. *Mon. Wea. Rev.,* **139**, 2008–2024, doi:10.1175/2010mwr3612.1.

Ancell, B. C., 2012: Examination of analysis and forecast errors of high-resolution assimilation, bias removal, and digital filter initialization with an ensemble Kalman filter. *Mon. Wea. Rev.*, **140**, 3992–4004, doi:10.1175/mwr-d-11-00319.1.

Ancell, B. C., E. Kashawlic, and J. L. Schroeder, 2015: Evaluation of wind forecasts and observation impacts from variational and ensemble data assimilation for wind energy applications. *Mon. Wea. Rev.*, **143**, 3230–3245, doi:10.1175/mwr-d-15-0001.1.

Arakawa, A., 2004: The cumulus parameterization problem: Past, present, and future. *J. Climate*, **17**, 2493–2525, doi:10.1175/1520-0442(2004)017,2493:RATCPP.2.0.CO;2.

Barker, D., and Coauthors, 2012: The weather research and forecasting model's community variational/ensemble data assimilation system: WRFDA. *Bull. Amer. Meteor. Soc.*, **93**, 831–843, doi:10.1175/bams-d-11-00167.1.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting,* **4**, 401–412, doi:10.1175/1520-0434(1989)004<0401:sfbotn>2.0.co;2.

Cattin, P., A. E. Gelfand, and J. Danes, 1983: A simple Bayesian procedure for estimation in a conjoint model. *J. Mark. Res.*, **20**, 29–35, doi:10.2307/3151409.

Chou, M. D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, Vol. 3, 85 pp.

Chu, P.-S., and X. Zhao, 2011: Bayesian analysis for extreme climatic events: A review. *Atmos. Res.,* **102**, 243–262, doi:10.1016/j.atmosres.2011.07.001.

Cotton, W. R., and Coauthors, 2003: RAMS 2001: Current status and future directions. *Meteor. Atmos. Phys.*, **82**, 5–29, doi:10.1007/s00703-001-0584-9.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, and R. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *J. Geophys. Res.*, **111**, D05308, doi:10.1029/2005jd006311.

Delle Monache, L., and Coauthors, 2008: A Kalman-filter bias correction method applied to deterministic, ensemble averaged and probabilistic forecasts of surface ozone. *Tellus*, **60B**, 238–249, doi:10.1111/j.1600-0889.2007.00332.x.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, and R. Stull, 2011: Kalman Filter and analog schemes to postprocess numerical weather predictions. *Mon. Wea. Rev.*, **139**, 3554–3570, doi:10.1175/2011mwr3653.1.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, doi:10.1175/mwr-d-12-00281.1

Djalalova, I., and Coauthors, 2010: Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006. *Atmos. Environ.*, **44**, 455–467, doi:10.1016/j.atmosenv.2009.11.007.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - II. Calibration and combination. *Tellus*, **57A**, 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Drusch, M., and P. Viterbo, 2007: Assimilation of screen-level variables in ECMWF's Integrated Forecast System: A study on the impact on the forecast quality and analyzed soil moisture. *Mon. Wea. Rev.*, **135**, 300–314, doi:10.1175/MWR3309.1.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, doi:10.1175/waf843.1.

Erickson, M. J., B. A. Colle, and J. J. Charney, 2012: Impact of Bias-correction type and conditional training on Bayesian model averaging over the northeast United States. *Wea. Forecasting*, **27**, 1449–1469, doi:10.1175/waf-d-11-00149.1.

Fountoukis, C., and A. Nenes, 2005: Continued development of a cloud droplet formation parameterization for global climate models. *J. Geophys. Res.*, **110**, D11212, doi:10.1029/2004jd005591.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190–202, doi:10.1175/2009mwr3046.1.

Frediani, M. E., J. P. Hacker, E.N. Anagnostou, and T. Hopson, 2016: Evaluation of PBL parameterizations for modeling surface wind speed during storms in the Northeast United States. *Wea. Forecasting*, **31**, 1511–1528, doi:10.1175/WAF-D-15-0139.1.

Gego, E., C. Hogrefe, G. Kallos, A. Voudouri, J. S. Irwin, and S. T. Rao, 2005: Examination of model predictions at different horizontal grid resolutions. *Environ. Fluid Mech.*, **5**, 63–85, doi:10.1007/s10652-005-0486-3.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:tuomos>2.0.co;2.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268, doi:10.1175/2008mwr2569.1.

Grell, G. A., and D. Devenyi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.*, **29**, 1693, doi:10.1029/2002gl015311.

Hacker, J. P., and D. L. Rife, 2007: A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Wea. Forecasting*, **22**, 1257–1273, doi:10.1175/2007waf2006102.1.

Hart, K. A., W. J. Steenburgh, D. J. Onton, and A. J. Siffert, 2004: An evaluation of mesoscale-model-based model output statistics (MOS) during the 2002 Olympic and

ParalympicWinter Games. *Wea. Forecasting*, **19**, 200–218, doi:10.1175/1520-0434(2004)019<0200:aeommo>2.0.co;2.

He, J., D. W. Wanik, B. M. Hartman, E. N. Anagnostou, M. Astitha, 2016: Nonparametric tree-based predictive modeling of storm damage to power distribution network. *Risk Anal.*, Accepted.

Homleid, M., 1995: Diurnal corrections of short-term surface temperature forecasts using Kalman filter. *Wea. Forecasting*, **10**, 689–707, doi:10.1175/1520-0434(1995)010<0689:dcosts>2.0.co;2.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, doi:10.1175/mwr3199.1.

Hu, X. M., J. W. Nielsen-Gammon, and F. Zhang, 2010: Evaluation of three planetary boundary layer schemes in the WRF Model. *J. Appl. Meteor. Climatol.*, **49**, 1831–1844, doi:10.1175/ 2010JAMC2432.1.

Idowu, O. S., and C. J. deW Rautenbach, 2009: Model Output Statistics to improve severe storms prediction over Western Sahel. *Atmos. Res.*, **93**, 419–425, doi:10.1016/j.atmosres.2008.10.035.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maxima and minima temperature, probability of precipitation, cloud amount, and surface wind. *Wea. Forecasting*, **5**, 128–138, doi:10.1175/1520-0434(1990)005<0128:nnbmgf>2.0.co;2.

Kang, D., R. Mathur, and S. T. Rao, 2010: Assessment of bias-adjusted PM2.5 air quality forecasts over the continental United States during 2007. *Geosci. Model Dev.*, **3**, 309–320, doi:10.5194/gmd-3-309-2010.

Kirtman, B. P., and Coauthors, 2014: The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/bams-d-12-00050.1.

Koster, R. D., and M. J. Suarez, 2001: Soil moisture memory in climate models. *J. Hydrometeor.*, **6**, 558–570, doi:10.1175/1525-7541(2001)002<0558:SMMICM>2.0.CO;2.

Krishnamurti, T. N., and Coauthors, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, doi:10.1126/science.285.5433.1548.

Krishnamurti, T. N., J. Sanjay, A. K. Mitra, and T. S. V. V. Kumar, 2004: Determination of forecast errors arising from different components of model physics and dynamics. *Mon. Wea. Rev.*, **132**, 2570–2594, doi:10.1175/mwr2785.1.

Kushta, J., G. Kallos, M. Astitha, S. Solomos, C. Spyrou, C. Mitsakou, and J. Lelieveld, 2014: Impact of natural aerosols on atmospheric radiation and consequent feedbacks with the meteorological and photochemical state of the atmosphere. *J. Geophys. Res. Atmos.*, **119**, 1463–1491, doi:10.1002/2013jd020714.

Libonati, R., I. Trigo, and C. C. Dacamara, 2008: Correction of 2 m-temperature forecasts using Kalman filtering technique. *Atmos. Res.*, **87**, 183–197, doi:10.1016/j.atmosres.2007.08.006.

Louka, P., G. Galanis, N. Siebert, G. Kariniotakis, P. Katsafados, I. Pytharoulis, and G. Kallos, 2008: Improvements in wind speed forecasts for wind power prediction purposes using Kalman filtering. *J. Wind Eng. Ind. Aerodyn.*, **96**, 2348–2362., doi:10.1016/j.jweia.2008.03.013.

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Wea. Forecasting*, **23**, 438–459, doi:10.1175/2007waf2006117.1.

Mao, Q., R. T. Mcnider, S. F. Mueller, and H.-M. H. Juang, 1999: An optimal model output calibration algorithm suitable for objective temperature forecasting. *Wea. Forecasting*, **14**, 190–202, doi:10.1175/1520-0434(1999)014<0190:aomoca>2.0.co;2.

McCollor, D., and R. Stull, 2008: Hydrometeorological accuracy enhancement via post-processing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144, doi:10.1175/2007waf2006107.1.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys.*, **20**, 851–875, doi:10.1029/rg020i004p00851.

Meyers, M. P., R. L. Walko, J. Y. Harrington, and W. R. Cotton, 1997: New RAMS cloud microphysics parameterization. Part II: The two-moment scheme. *Atmos. Res.*, **45**, 3–39, doi:10.1016/s0169-8095(97)00018-5.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16663–16682, doi:10.1029/97jd00237.

Müller, M. D., 2011: Effects of model resolution and statistical postprocessing on shelter temperature and wind forecasts. *J. Appl. Meteor. Climatol.*, **50**, 1627–1636, doi:10.1175/2011jamc2615.1.

Murphy, A. H., 1988: Skill score based on the mean square error and their relationship to the correlation coefficient. *Mon. Wea. Rev.*, 116, 2417–2424.

Nenes, A., and J. H. Seinfeld, 2003: Parameterization of cloud droplet formation in global climate models. *J. Geophys. Res.*, **108**, 4415, doi:10.1029/2002jd002911.

Nielsen-Gammon, J.W., X.-M. Hu, F. Zhang, and J. E. Pleim, 2010: Evaluation of planetary boundary layer scheme sensitivities for the purpose of parameter estimation. *Mon. Wea. Rev.*, **138**, 3400–3417, doi:10.1175/2010MWR3292.1.

NCEP/NWS/NOAA/DOC, 2000. Updated daily. NCEP FNL Operational Model Global Tropospheric Analyses, continuing from July 1999. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://dx.doi.org/10.5065/D6M043C6.

NCEP/NWS/NOAA/DOC, 2007. NCEP Global Forecast System (GFS) Analyses and Forecasts. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds084.6/.

O'Hagan, A., and J. J. Forster, 2004: *Kendall's Advanced Theory of Statistics, Volume 2B: Bayesian Inference*. Arnold, 496 pp.

Palmer, T. N., F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell, 2008: Toward seamless prediction: Calibration of climate change projections using seasonal forecasts. *Bull. Amer. Meteor. Soc.*, **89**, 459–470, doi:10.1175/bams-89-4-459.

Papadopoulos, A., E. Serpetzoglou, and E. N. Anagnostou, 2008: Improving NWP through radar rainfall-driven land surface parameters: A case study on convective precipitation forecasting. *Adv. Water Res.*, **31**, 1456–1469, doi:10.1016/j.advwatres.2008.02.001.

Pielke, R. A., and Coauthors, 1992: A comprehensive meteorological modeling system–RAMS. *Meteor. Atmos. Phys.*, **49**, 69–91, doi:10.1007/bf01025401.

Pleim, J. E., 2007: A combined local and nonlocal closure model for the atmospheric boundary layer. Part II: Application and evaluation in a mesoscale meteorological model. *J. Appl.Meteor. Climatol.*, **46**, 1396–1409, doi:10.1175/JAM2534.1.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/mwr2906.1.

Rincon, A., O. Jorba, and J. M. Baldasano, 2010: Development of a short-term irradiance prediction system using post-processing tools on WRF-ARW meteorological forecasts in Spain. *Extended Abstracts, European Conf. on Applied Meteorology*, Vol. 7, Zurich, Switzerland, European Meteorological Society, EMS2010-406.

Roberts, N. M., 2003: Results from high-resolution simulations of convective events. Met Office Tech. Rep. 402, 47 pp.

Roeger, C., R. Stull, D. Mcclung, J. Hacker, X. Deng, and H. Modzelewski, 2003: Verification of mesoscale numerical weather forecast in mountainous terrain for application to avalanche prediction. *Wea. Forecasting*, **18**, 1140–1160, doi:10.1175/1520-0434(2003)018<1140:vomnwf>2.0.co;2.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi:10.1175/2009mwr2924.1.

Serpetzoglou, E., E. N. Anagnostou, A. Papadopoulos, E. I. Nikolopoulos, and V. Maggioni, 2010: Error propagation of remote sensing rainfall estimates in soil moisture prediction from a land surface model. *J. Hydrometeor.*, **11**, 705–720, doi:10.1175/2009JHM1166.1.

Simmons, A., S. Uppala, D. Dee, and S. Kobayashi, 2007: ERA-Interim: New ECMWF reanalysis products from 1989 onwards. ECMWF Newsletter, No. 110, ECMWF, Reading, United Kingdom, 25–35. [Available online at http://www.ecmwf.int/publications/newsletters/pdf/110_rev.pdf.]

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-4751STR, 113 pp. [Available online at http://www.mmm.ucar.edu/wrf/users/docs/arw_v3.pdf.]

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/mwr3441.1.

Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.*, **105**, 25–35, doi:10.1198/jasa.2009.ap08615.

Solomos, S., G. Kallos, J. Kushta, M. Astitha, C. Tremback, A. Nenes, and Z. Levin, 2011: An integrated modeling study on the effects of mineral dust and sea salt particles on clouds and precipitation. *Atmos. Chem. Phys.*, **11**, 873–892, doi:10.5194/acp-11-873-2011.

Sorensen, D., and D. Gianola, 2002: *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics*. Springer, 740 pp.

Speer, M. S., L. M. Leslie, and L. Qi, 2003: Numerical prediction of severe convection: comparison with operational forecasts. *Meteor. Appl.*, **10**, 11–19, doi:10.1017/s1350482703005024.

Stensrud, D. J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524, doi:10.1175/1520-0493(2003)131<2510:sepomt>2.0.co;2.

Stensrud, D. J., and N. Yussouf, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12**, 217, doi:10.1017/s135048270500174x.

Steppeler, J., G. Doms, U. Schättler, H. W. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteor. Atmos. Phys.*, **82**, 75–96, doi:10.1007/s00703-001-0592-9.

Sweeney, C. P., P. Lynch, and P. Nolan, 2011: Reducing errors of wind speed forecasts by an optimal combination of post-processing methods. *Meteor. Appl.*, **20**, 32–40, doi:10.1002/met.294.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.*, **106**, 7183–7192,doi:10.1029/2000jd900719.

Tewari, M., F. Chen, W. Wang, J. Dudhia, M. A. LeMone, K. Mitchell, M. Ek, G. Gayno, J. Wegiel, and R. H. Cuenca, 2004: Implementation and verification of the unified NOAH land surface model in the WRF model. *20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction*, pp. 11–15, Seattle, USA, American Meteorological Society.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, doi:10.1175/2008mwr2387.1.

Walko, R. L., W. Cotton, M. Meyers, and J. Harrington, 1995: New RAMS cloud microphysics parameterization part I: the single-moment scheme. *Atmos. Res.*, **38**, 29–62, doi:10.1016/0169-8095(94)00087-t.

Walko, R. L., and Coauthors, 2000: Coupled atmosphere–biophysics–hydrology models for environmental modeling. *J. Appl. Meteor.*, **39**, 931–944, doi:10.1175/1520-0450(2000)039<0931:cabhmf>2.0.CO;2.

Walter, G., T. Augustin, and A. Peters, 2007: Linear regression analysis under sets of conjugate priors. *ISIPTA '07: Proc. 5th Int. Symp. on Imprecise Probabilities and Their Applications*, Vol. 7, 445–455, Prague, Czech Republic.

Walter, G., and T. Augustin, 2010: Bayesian linear regression–different conjugate models and their (in) sensitivity to prior-data conflict. *Statistical Modelling and Regression Structures*, Kneib, T., and G. Tutz, Physica-Verlag HD, 59–78.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117, doi:10.1175/mwr-d-12-00141.1.

Wanik, D.W., E. Anagnostou, B.M. Hartman, M.E. Frediani, M. Astitha, 2015: Storm outage modeling for an electric distribution network in Northeastern USA. *Nat. Hazards*, **79**, 1359–1384, doi 10.1007/s11069-015-1908-2.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2009: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Mon. Wea. Rev.*, **137**, 1460–1479, doi:10.1175/2008mwr2773.1.

Weisberg, S. 2005: *Applied linear regression*. John Wiley & Sons, 310 pp.

Wilczak, J. M., S. A. McKeen, I. Djalalova, and G. Grell, 2006: Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during summer of 2004. *J. Geophys. Res.*, **111**, D23S28, doi:10.1029/2006JD007598.

Wilks, D., S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, doi:10.1175/mwr3402.1.

Wilson, L. J., and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222, doi:10.1175/1520-0434(2002)017<0206:tcumos>2.0.co;2.

Wilson, L. J., and M. Vallée, 2003: The Canadian updateable model output statistics (UMOS) system: Validation against perfect prog. *Wea. Forecasting*, **18**, 288–302, doi:10.1175/1520-0434(2003)018<0288:tcumos>2.0.co;2.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 1364–1385, doi:10.1175/mwr3347.1.

Zellner, A., 1996: An Introduction to Bayesian Inference in Econometrics. John Wiley & Sons, 431 pp.

FIGURE LIST

TABLE 1. WRF and ICLAMS configuration.

	WRF	ICLAMS
Grid structure	3 grids Horizontal: 18-6-2 km Vertical: 27 levels (P_{top} = 50 hPa)	3 grids Horizontal: 18-6-2 km Vertical: 50 levels (P_{top} = 60 hPa)
Horizontal grid scheme	Arakawa C grid	Arakawa C grid
Nesting	2-way nesting	2-way nesting
Initial Conditions	NCEP GFS ($1^{\circ} \times 1^{\circ}$, 6-hour)	NCEP FNL ($1^{\circ} \times 1^{\circ}$, 6-hour)
Cumulus scheme (per grid)	Grell 3D scheme (Grell and Devenyi 2002) on the parent and second grids; no parameterization on the third grid	Kain-Fritsch cummulus parameterization on the parent and second grids; no parameterization on the third grid
Cloud microphysics	Thompson et al. (2008) scheme	Warm rain processes; Five ice condensate species; Two-moment bulk scheme (Walko et al. 1995; Meyers et al. 1997); Explicit cloud droplet activation scheme (Nenes and Seinfeld, 2003; Fountoukis and Nenes, 2005) with prescribed aerosols.
Planetary boundary layer	Yonsei scheme (Hong et al. 2006)	Mellor- Yamada scheme (1982)
Boundary Conditions	SST (NCEP GFS); topography (USGS GTOPO30, 30”); land cover (USGS, 30’’); soil texture (FAO, 5’; North-America STATSGO, 30”)	SST daily; NDVI (USGS, 30’’); topography (NASA SRTM90 v4.1, 3’’); land cover (USGS OGE, 30’’); soil texture (FAO, 2’)
Radiation	Goddard for short wave radiation (Chou and Suarez 1994); RRTM for long wave radiation (Mlawer et al. 1997)	RRTM for short/long wave radiation (Mlawer et al. 1997)
Land surface	WRF NOAH (Tewari et al. 2004)	LEAF-3 (Walko et al., 2000)

TABLE 2. Storm type and date (the duration of all simulations was 61h).

Storms	Model start date and time	Storm type
Nov 5th 2004	4 Nov 2004 1800 UTC	Wind storm
Dec 1st 2004	30 Nov 2004 1200 UTC	Wind storm
Apr 1-3rd 2005	1 Apr 2005 1800 UTC	Rain/Wind storm
Oct 16th 2005	15 Oct 2005 1800 UTC	Wind storm
Oct 24-25th 2005	24 Oct 2005 1800 UTC	Nor’easter
Jan 14-15th 2006	14 Jan 2006 0000 UTC	Rain/Snow/Wind storm
Jan 18th 2006	17 Jan 2006 0000 UTC	High wind storm
Feb 17th 2006	16 Feb 2006 1800 UTC	High wind storm
Apr 15-16th 2007	14 Apr 2007 1800 UTC	Nor’easter
Jan 6-8th 2009	7 Jan 2009 0000 UTC	Ice Storm/Wind storm
Mar 12-15th 2010	13 Mar 2010 0000 UTC	Heavy rain/High wind storm
Dec 26-27th 2010	25 Dec 2010 1800 UTC	Blizzard
Jan 11-12th 2011	11 Jan 2011 1800 UTC	Heavy Snow storm
Aug 28th 2011	28 Aug 2011 0000 UTC	Hurricane (Irene)
Oct 29th 2012	28 Oct 2012 1800 UTC	Hurricane (Sandy)
Nov 7th 2012	7 Nov 2012 0600 UTC	Nor’easter
Feb 8-9th 2013	8 Feb 2013 0000 UTC	Blizzard

(Source: NOAA, Significant Weather Events Archive; http://www.erh.noaa.gov/okx/stormtotals.html)

TABLE 3. Number of possible storm sequence combinations for the training datasets.

Number of storms (r)	Number of r-combinations
1	13
2	78
3	286
4	715
5	1287
6	1716
7	1716
8	1287
9	715
10	286
11	78
12	13
13	1
All possible combinations	8191

TABLE 4. Skill score (%) evaluated by RMSE and R^2 of WRF_{SLR}, $ICLAMS_{SLR}$ and BLR with N_e = 13 (N_e: Number of storms in the training dataset).

Storm	SS_{RMSE} (%)				SS_{R^2} (%)			
	WRF_{SLR} vs. WRF_{raw}	BLR vs. WRF_{raw}	$ICLAMS_{SLR}$ vs. $ICLAMS_{raw}$	BLR vs. $ICLAMS_{raw}$	WRF_{SLR} vs. WRF_{raw}	BLR vs. WRF_{raw}	$ICLAMS_{SLR}$ vs. $ICLAMS_{raw}$	BLR vs. $ICLAMS_{raw}$
Irene	20.4	24.6	9.2	17.1	12.0	15.7	8.0	14.5
Sandy	16.6	22.0	15.1	18.3	21.0	32.7	16.7	23.7
7 Nov. 2012	24.4	31.9	21.9	21.7	27.0	47.3	25.5	25.5
8-9 Feb. 2013	22.9	31.3	15.0	18.0	20.0	33.5	18.8	22.8

TABLE 5. Normalized RMSE and R^2 by the relevant metrics for BLR with N_e = 13.

Storm	RMSE normalized by BLR				R^2 normalized by BLR			
	$\frac{WRF_{raw}}{BLR}$	$\frac{ICLAMS_{raw}}{BLR}$	$\frac{WRF_{SLR}}{BLR}$	$\frac{ICLAMS_{SLR}}{BLR}$	$\frac{WRF_{raw}}{BLR}$	$\frac{ICLAMS_{raw}}{BLR}$	$\frac{WRF_{SLR}}{BLR}$	$\frac{ICLAMS_{SLR}}{BLR}$
Irene	1.33	1.21	1.06	1.10	0.86	0.87	0.97	0.94
Sandy	1.28	1.22	1.07	1.04	0.75	0.81	0.91	0.94
7 Nov. 2012	1.47	1.28	1.11	1.00	0.68	0.80	0.86	1.00
8-9 Feb. 2013	1.46	1.22	1.12	1.04	0.75	0.81	0.90	0.97

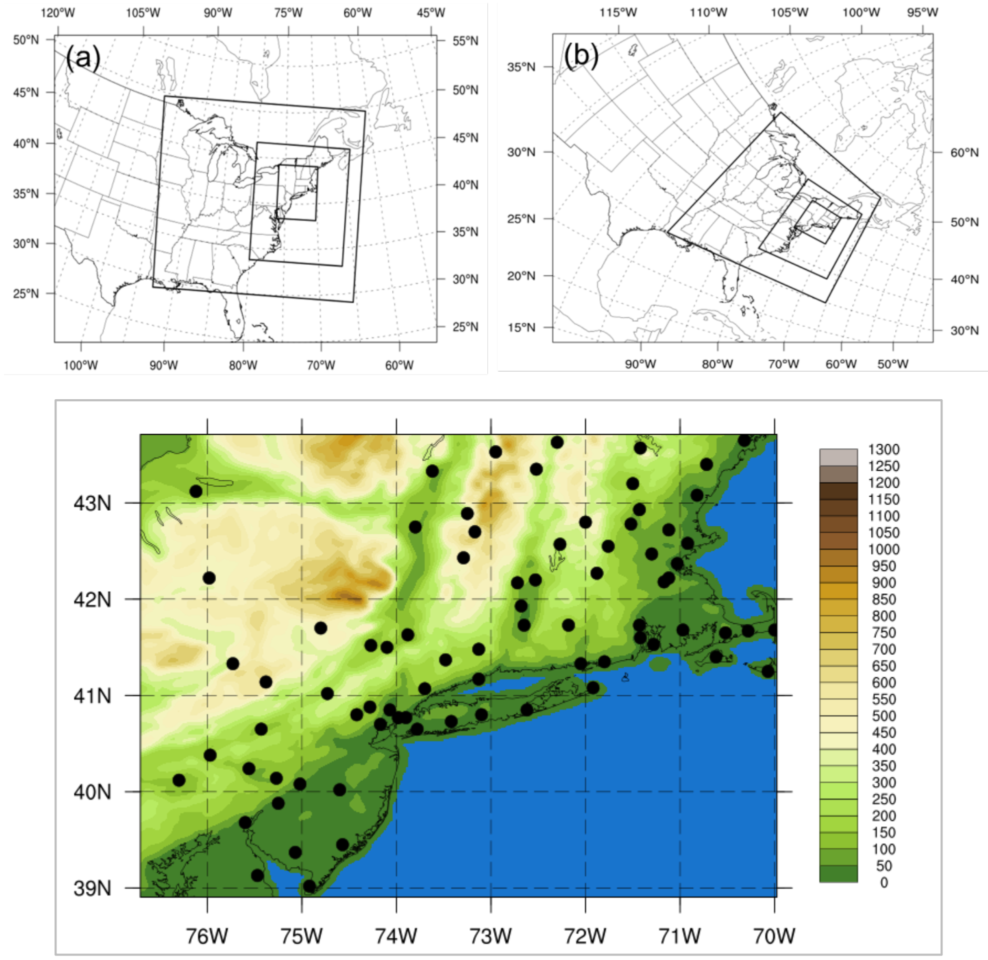

Figure 1. Model domains: (a) WRF and (b) RAMS/ICLAMS; (c) NCEP/NWS/NOAA stations over the NE U.S. (black circles) and elevation (m).

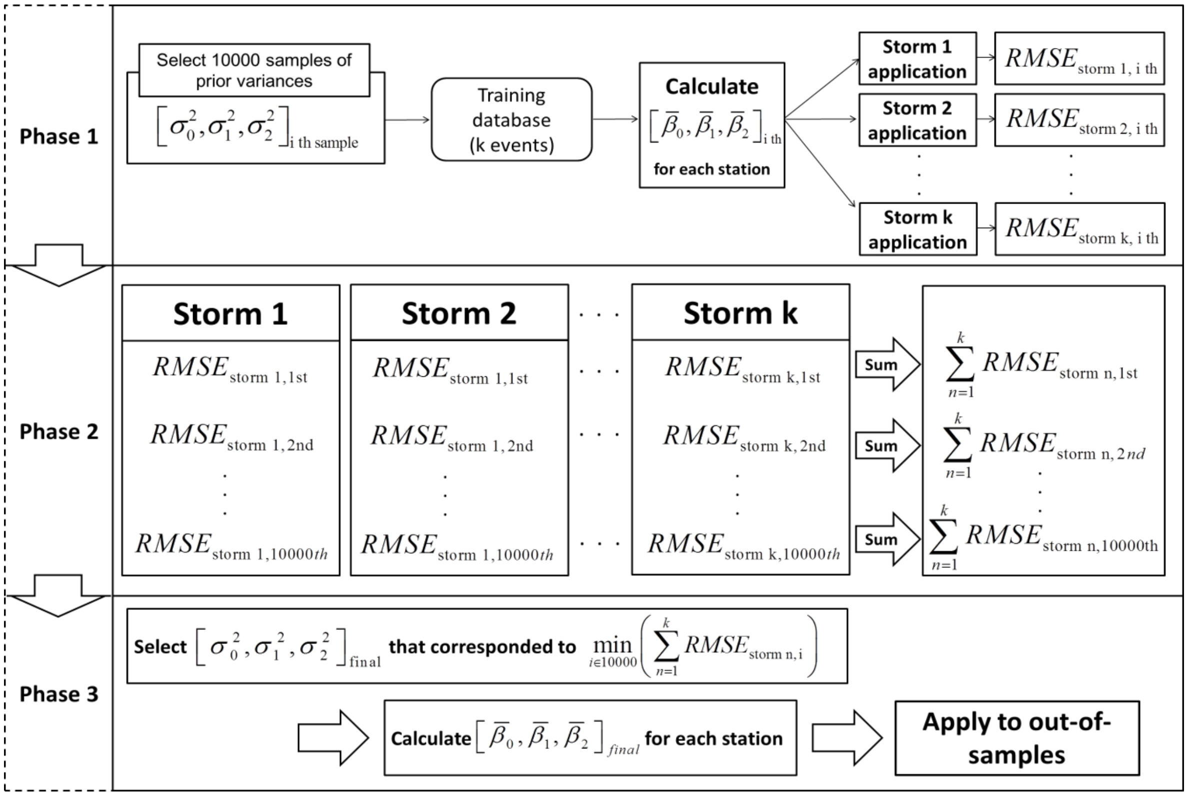

Figure 2. A schematic diagram of the Bayesian Linear Regression (BLR) approach.

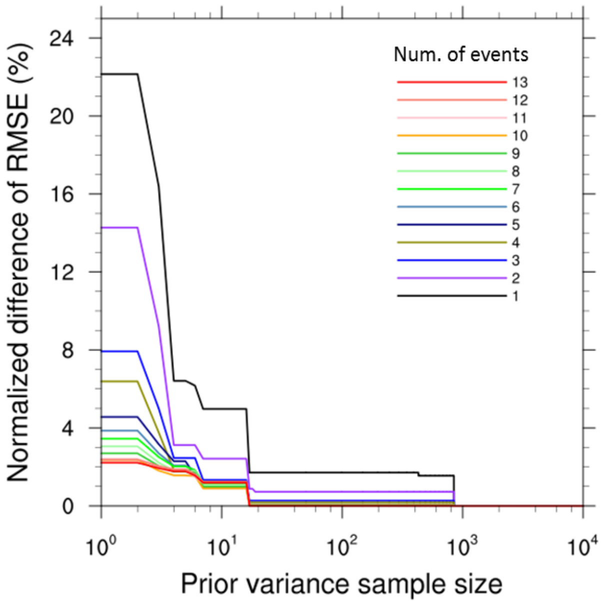

Figure 3. RMSE normalized difference using different sample sizes for the BLR training datasets.

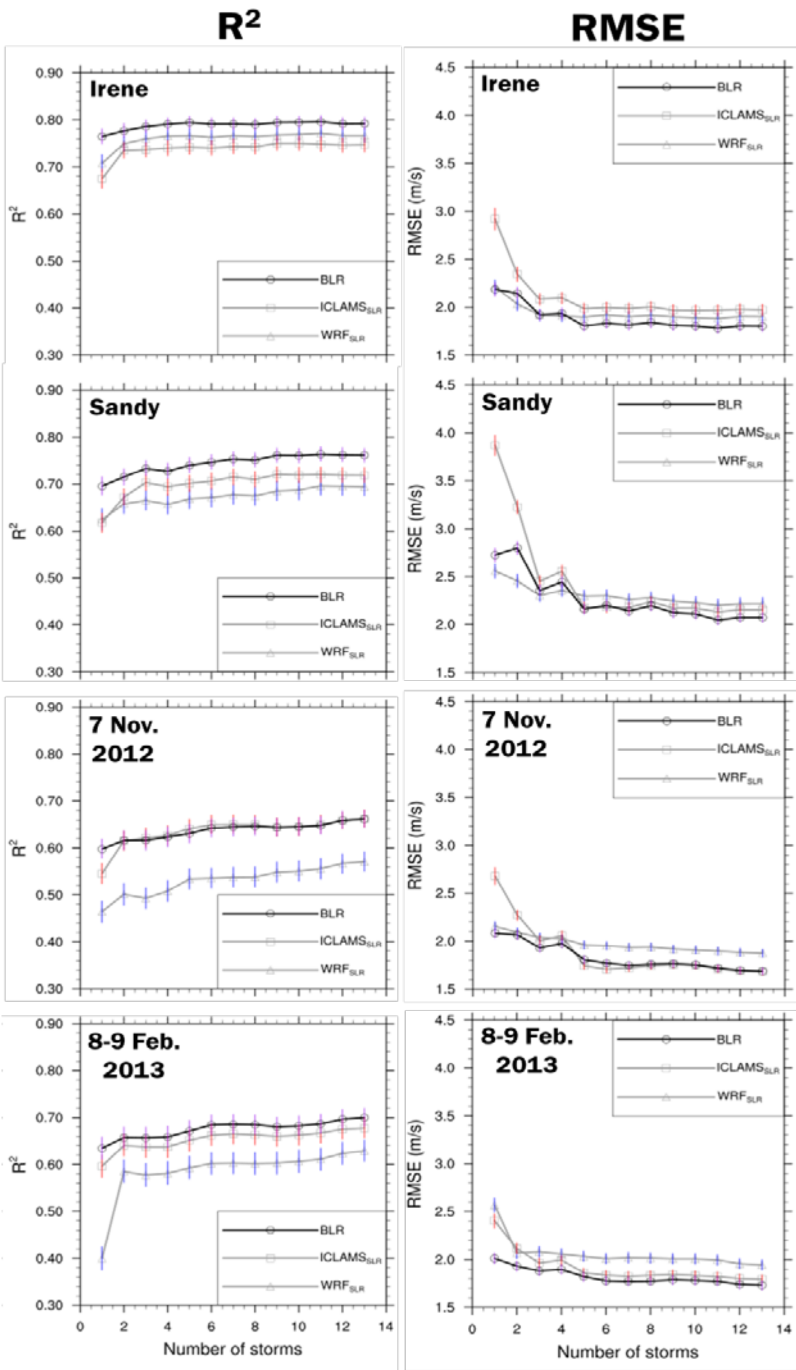

Figure 4. Chronological storm sequence experiment: R^2 and RMSE (m/s) variation by increasing the number of training storms for WRF_{SLR} (triangles), $ICLAMS_{SLR}$ (squares), BLR (circles). The 95% bootstrap confidence intervals are indicated by the error bars for WRF_{SLR} (blue), $ICLAMS_{SLR}$ (red), BLR (purple).

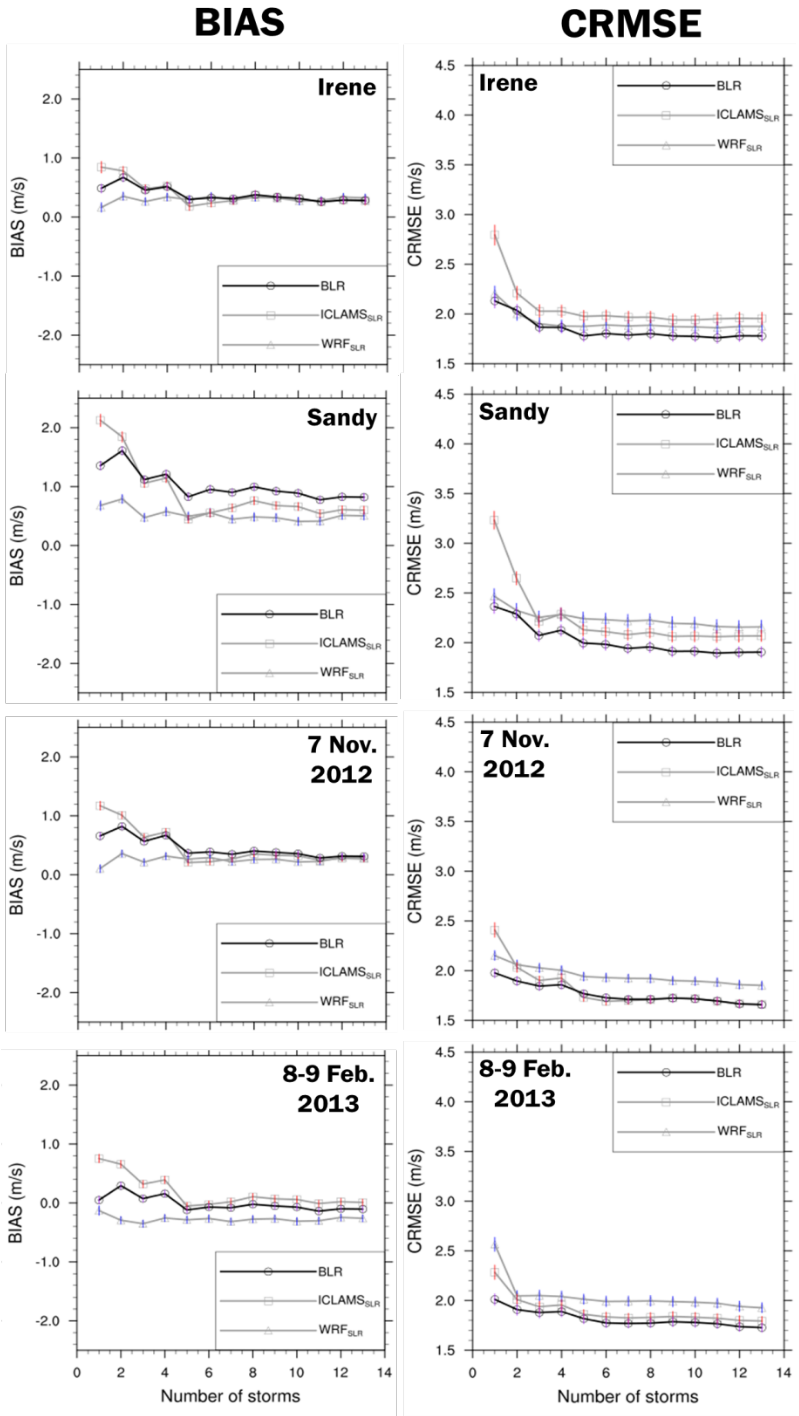

Figure 5. As in Fig. 4, but for BIAS (m/s) and CRMSE (m/s).

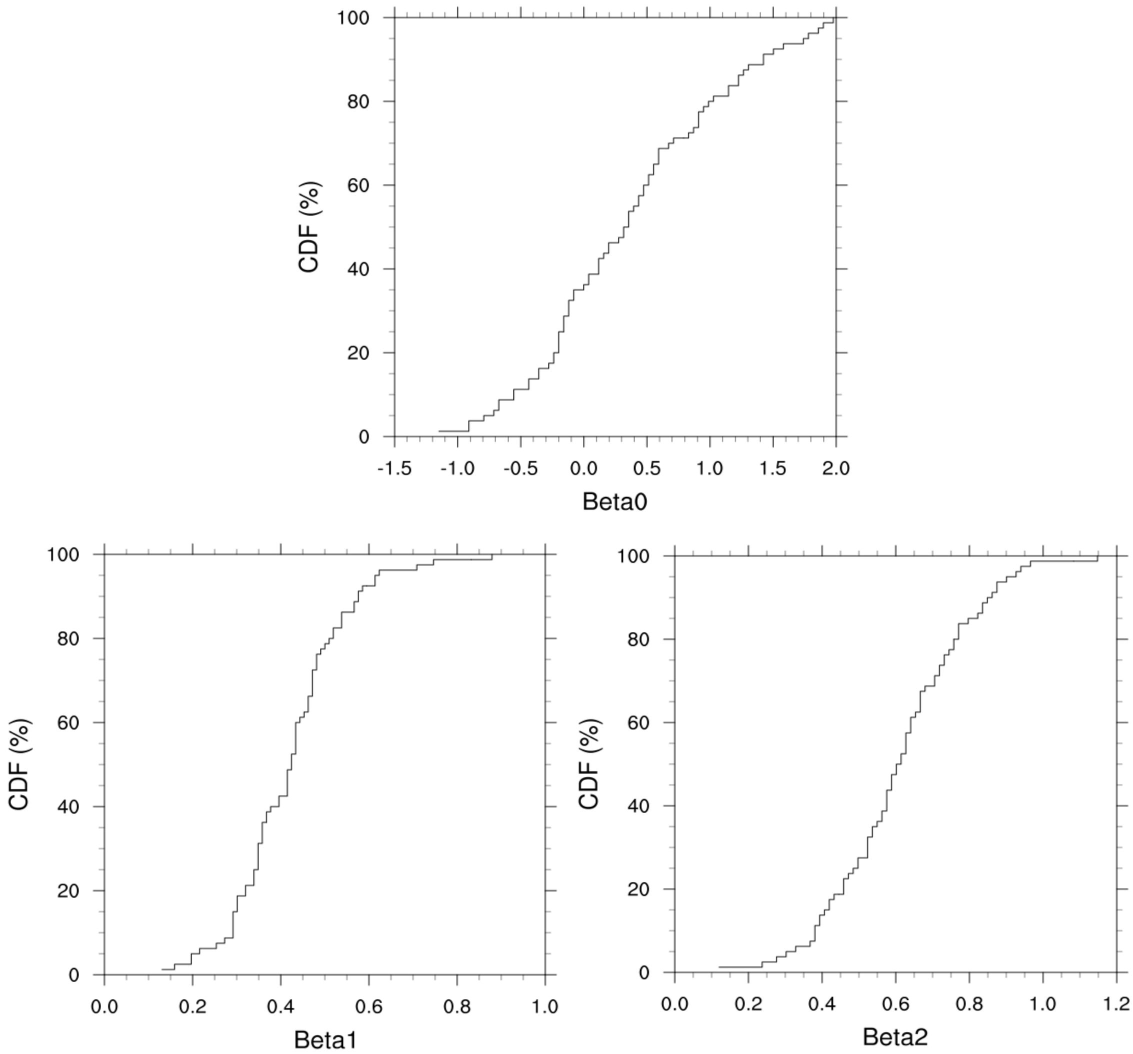

Figure 6. Cumulative distribution function (CDF) of the beta coefficients at 80 stations when including 13 storms in the training dataset (BLR = β0+β1·WRF+β2·ICLAMS).

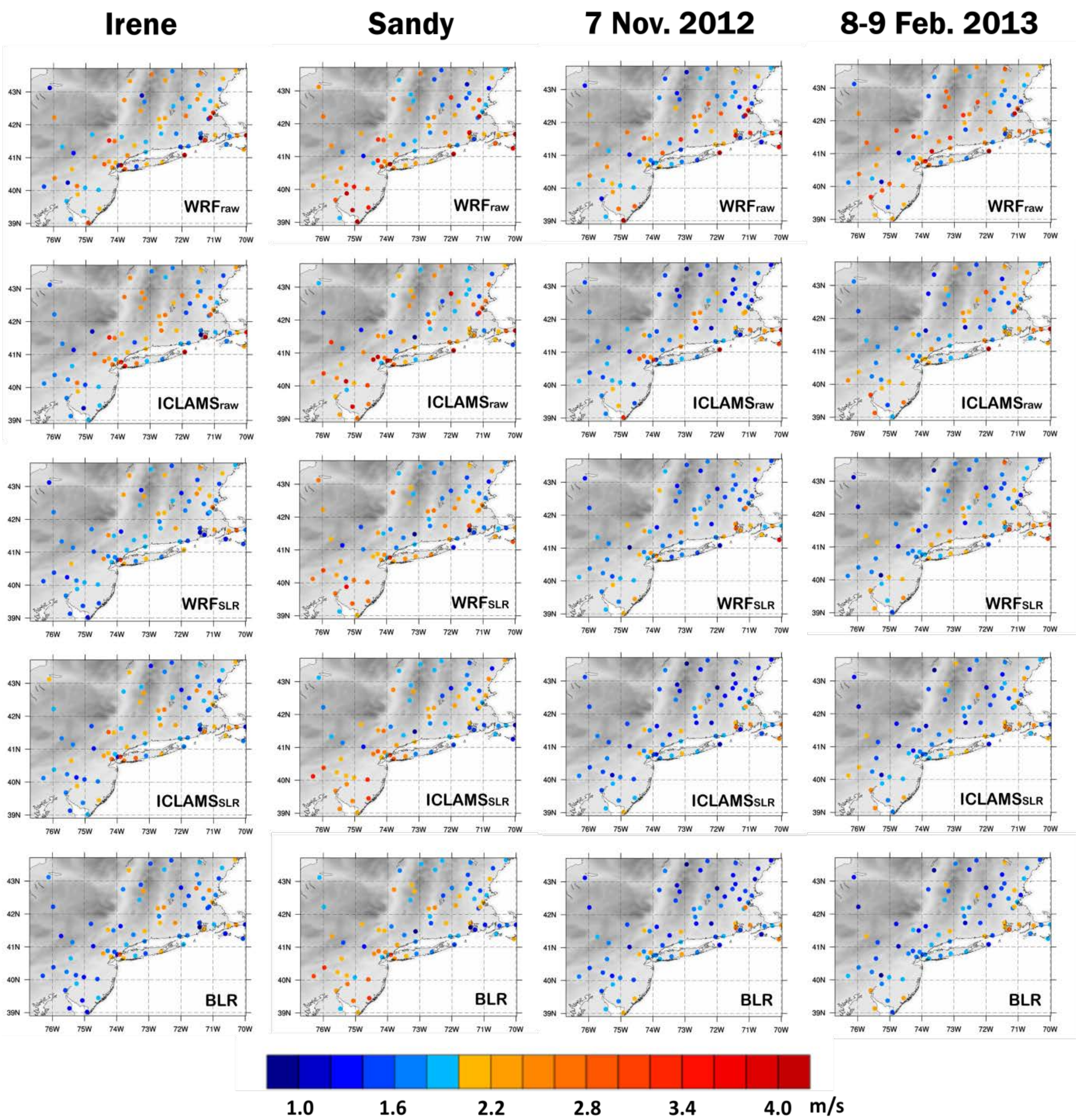

Figure 7. Spatial distribution of RMSE for the chronologically-ordered training dataset application.

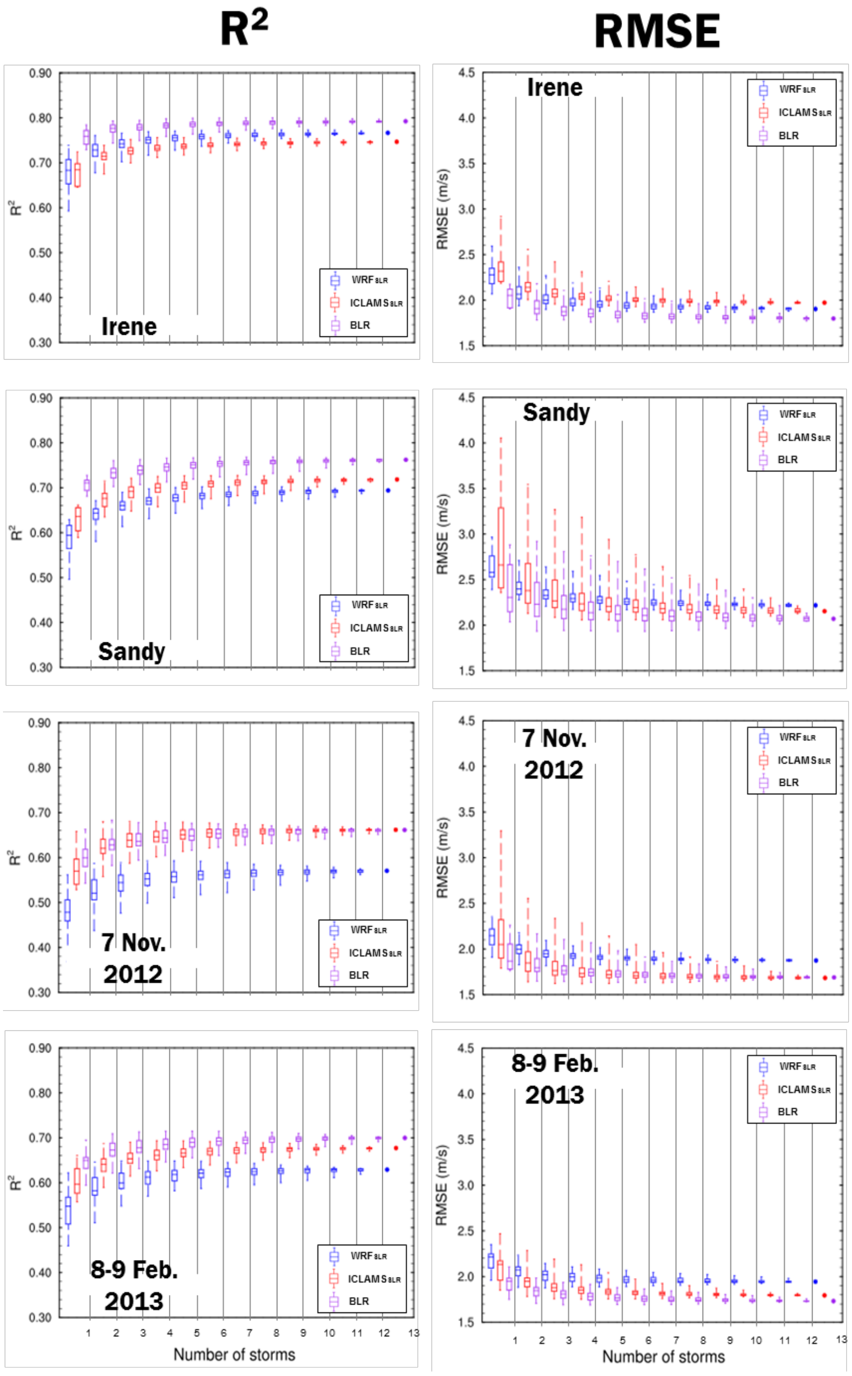

Figure 8. Randomized storm sequence experiment: R^2 and RMSE (m/s) spread behavior (bar: median, box: interquartile range, whiskers: range, and error bars: minimum and maximum) related to the number of storms in the training dataset for WRF_{SLR} (blue), $ICLAMS^{SLR}$ (red), BLR (purple).

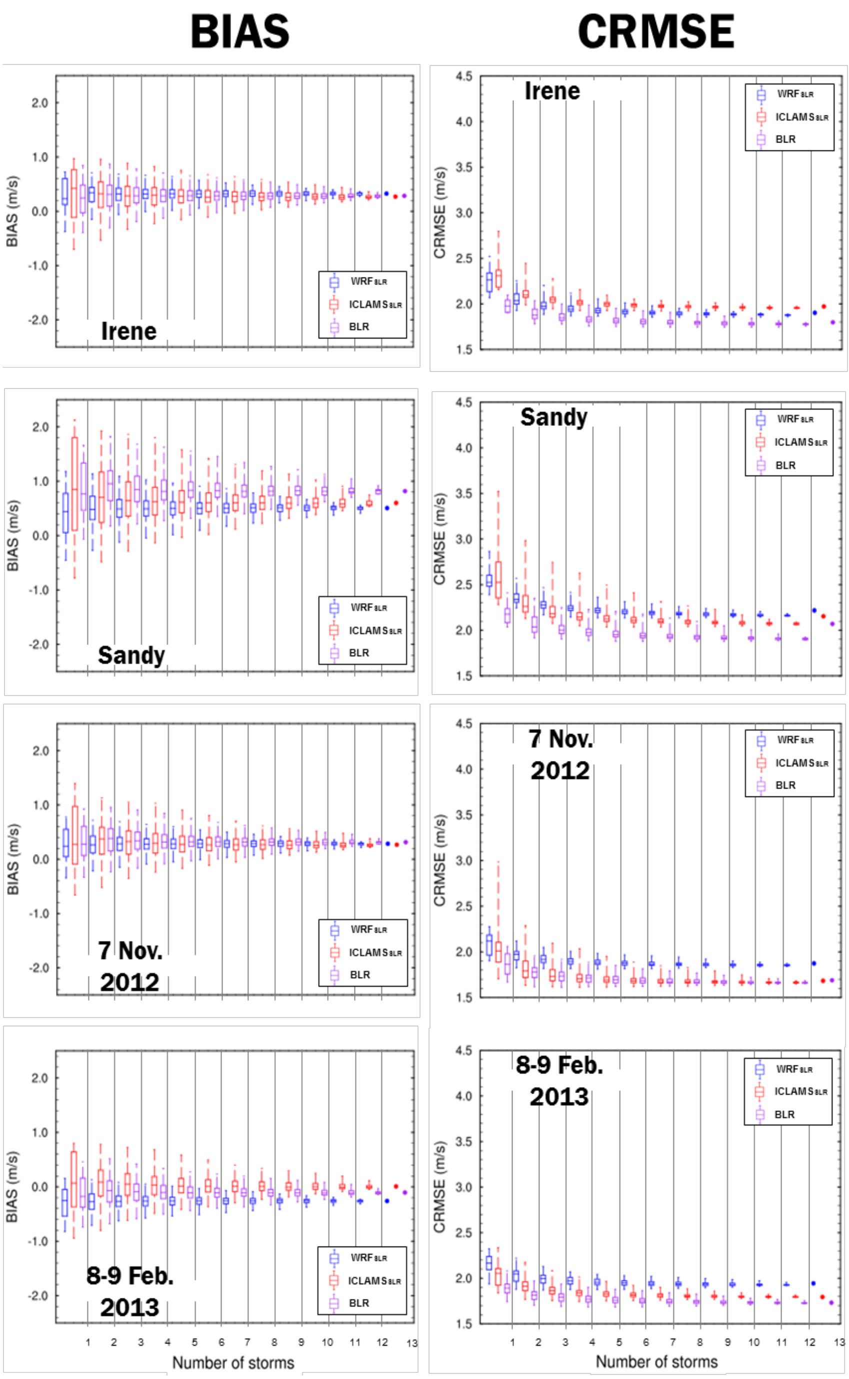

Figure 9. As in Fig. 8, but for BIAS (m/s) and CRMSE (m/s).