Updating Short-Term Probabilistic Weather Forecasts of Continuous Variables Using Recent Observations

THOMAS N. NIPEN, GREG WEST, AND ROLAND B. STULL

University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 16 February 2011, in final form 5 June 2011)

ABSTRACT

A statistical postprocessing method for improving probabilistic forecasts of continuous weather variables, given recent observations, is presented. The method updates an existing probabilistic forecast by incorporating observations reported in the intermediary time since model initialization. As such, this method provides updated short-range probabilistic forecasts at an extremely low computational cost. The method models the time sequence of cumulative distribution function (CDF) values corresponding to the observation as a first-order Markov process. Verifying CDF values are highly correlated in time, and their changes in time are modeled probabilistically by a transition function. The effect of the method is that the spread of the probabilistic forecasts for the first few hours after an observation has been made is considerably narrower than the original forecast. The updated probability distributions widen back toward the original forecast for forecast times far in the future as the effect of the recent observation diminishes. The method is tested on probabilistic forecasts produced by an operational ensemble forecasting system. The method improves the ignorance score and the continuous ranked probability score of the probabilistic forecasts significantly for the first few hours after an observation has been made. The mean absolute error of the median of the probability distribution is also shown to be improved.

1. Introduction

Correctly predicting forecast uncertainty can bring significant economic benefits to many decision makers (AMS 2008). Unlike a deterministic forecast, which supplies only the expected weather outcome, a probabilistic forecast gives the likelihood of occurrence of all outcomes. Decisions are based on combining the relative risks of various weather outcomes with the costs and losses corresponding to those outcomes. Thus, probabilistic forecasts are naturally preferred for economic decision making.

Let $f_t(x)$ be the forecasted probability density function (PDF) of a continuous meteorological variable X (such as temperature) valid for time t. One can generate $f_t(x)$ from an ensemble of numerical weather prediction (NWP) models by using methods such as Bayesian model averaging (Raftery et al. 2005), the binned probability ensemble technique (Anderson 1996), the method of moments (Jewson et al. 2005), or local quantile regression (Bremnes 2004).

Let $F_t(x)$ denote the forecasted cumulative distribution function (CDF) given by

$$F_t(x) = \int_{-\infty}^{x} f_t(s)\, ds. \tag{1}$$

In addition, let x_t denote the observed state of X at time t. Let p_t denote the CDF value corresponding to the observed state:

$$p_t = F_t(x_t). \tag{2}$$

Often, p_t is called the probability integral transform value (PIT value) corresponding to the observation.

We will assume an operational ensemble forecasting system initialized at time $t = 0$ that gives hourly forecasts out to time $t = T$. At times t, where $0 \le t \le T$, hourly observations from observing stations are made available, but the models do not incorporate these observations until the next forecast cycle starts.

Figure 1a shows a sample temperature CDF forecast for a single location produced from an ensemble. At

Corresponding author address: Thomas Nipen, Dept. of Earth and Ocean Sciences, 6339 Stores Rd., Vancouver BC V6T 1Z4, Canada.
E-mail: tnipen@eos.ubc.ca

DOI: 10.1175/WAF-D-11-00022.1

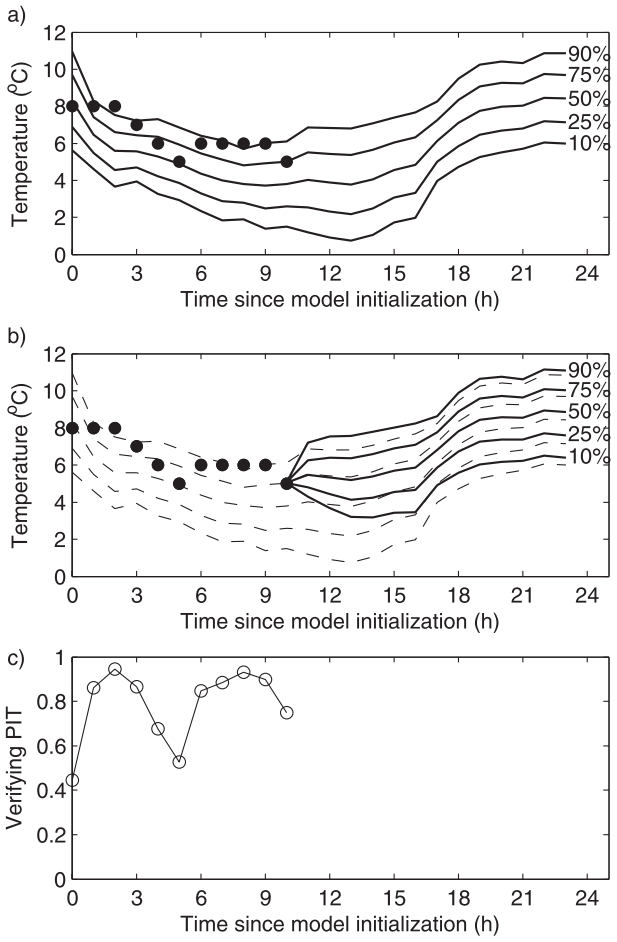

FIG. 1. (a) A sample probabilistic temperature forecast initialized at 0000 UTC. Forecasted cumulative probability values are shown by lines. Observations are shown by solid dots. (b) The updated probabilistic forecast (solid lines) based on the most recent observation. The original forecast is shown by dashed lines. (c) The probability integral transform values of the original forecast corresponding to the observations.

the time the figure was produced, observations up to 1000 UTC were available. What is clear from the figure is that the CDF value that the observation verifies on (PIT value) is highly correlated in time (Fig. 1c). Given that the most recent PIT value (at 1000 UTC) is 0.75, the next PIT value (at 1100 UTC) will likely be near 0.75.

The probability distribution can therefore be refined to take into account this new information that was not available at the time the model was initialized. The effects of the most recent observation will diminish for longer lead times. The updated probability distribution will therefore be narrow near the time of the observation and widen back to the original distribution for times in the future (Fig. 1b).

The goal of this paper is to present a method for producing an updated probabilistic forecast $\hat{F}_t(x)$ by mapping the original CDF $F_t(x)$ by a function Φ as follows:

$$\hat{F}_t(x) = \Phi[F_t(x)]. \tag{3}$$

The mapping will concentrate $\hat{F}$ in a narrower range with the hope of improving short-term verification scores. End users in need of rapidly updating probabilistic short-term forecasts at very low computational costs can benefit from this update method.

Postprocessing weather forecasts is commonly done to increase the correspondence between forecasts and observations. For deterministic forecasts, methods such as model output statistics (Glahn and Lowry 1972), Kalman filtering (Homleid 1995), and analog methods (Delle Monache et al. 2011) are commonly used to reduce forecast error. On the other hand, methods such as ensemble calibration (Hamill and Colucci 1998) and Bayesian model averaging (Raftery et al. 2005) can be used to improve probabilistic forecasts from an ensemble of deterministic forecasts. The method presented here also aims to improve probabilistic forecasts, but differs in that it is only invoked once observations are available after the raw forecasts are created. It is therefore of most use for operational short-term forecasts.

This paper is organized as follows: the method for updating probabilistic forecasts is presented in section 2, the dataset and verification metric used for testing the method is described in section 3, the performance of the method is evaluated in section 4, and conclusions are drawn in section 5.

2. Method

Assume that for a given forecast day, $T + 1$ hourly probabilistic forecasts $F_t(x)$ (where $0 \le t \le T$) are produced. Let t_{obs} denote the time at which the most recent observation was made. This observation is then used to update all hourly forecasts that are still in the future (i.e., where $t_{\text{obs}} < t \le T$).

The probabilistic forecast n hours after t_{obs}, that is for time $t = t_{\text{obs}} + n$, can be updated according to

$$\hat{F}_{t_{\text{obs}}+n}(x) = \Phi_n[F_{t_{\text{obs}}+n}(x)], \tag{4}$$

where $\Phi_n(p)$ will in general be different for each value of n and can be constructed based on forecast and observation data prior to the time t_{obs}. Here, $\Phi_n(p)$ is the probability function that the verifying PIT value of the original forecast will be less than p.

Combining Eqs. (1) and (4) and using the chain rule gives the following for the updated PDF:

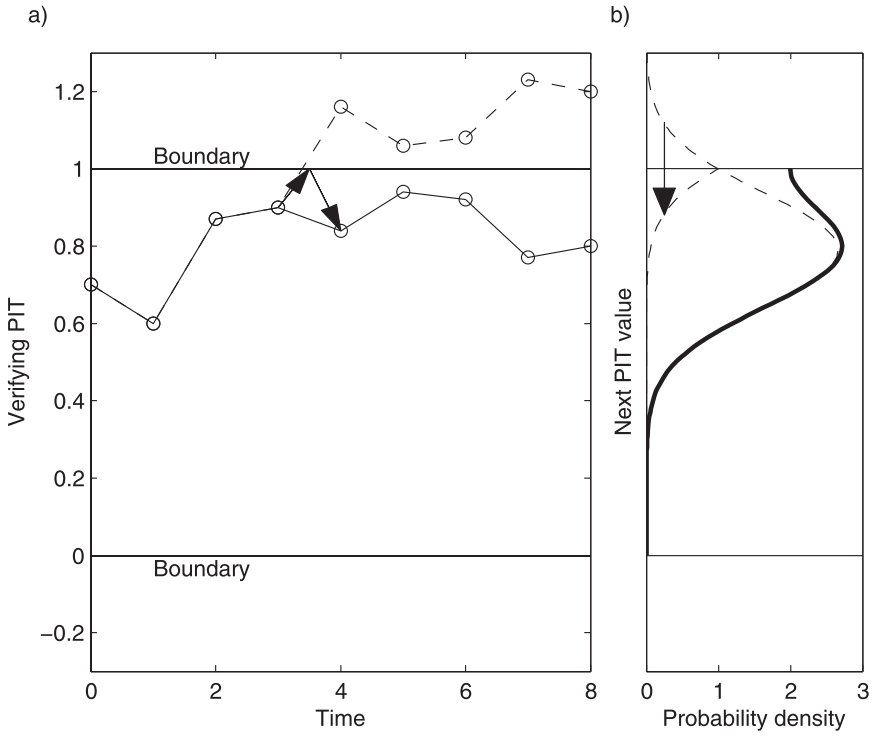

FIG. 2. (a) A hypothetical time series of verifying PIT values (solid line). Mirror barriers at 0 and 1 reflect any steps back into the domain. The dashed line shows the PIT time series without reflections. The transition from time 3 to 4 involves a reflection across 1 as shown by the arrows. (b) The PDF (thick solid line) of the PIT value for time 9, given that the PIT value at time 8 was 0.80. The dashed line shows the probability of the Gaussian distribution that has been reflected back into the domain.

$$\hat{f}_{t_{\text{obs}}+n}(x) = \Psi_n[F_{t_{\text{obs}}+n}(x)]f_{t_{\text{obs}}+n}(x), \tag{5}$$

where $\Psi_n(p)$ is the derivative of $\Phi_n(p)$ and acts as an amplification factor for the original PDF. We note that $\Psi_n(p)$ increases probability density in regions where the PIT value is more likely to occur given the recent observation. That is, $\Psi_n(p)$ is also the probability density of p being the verifying PIT value of the original forecast.

a. PIT values as a random walk in time

We model the time sequence of verifying PIT values within one forecast cycle as a random walk in time. Mirror barriers at 0 and 1 are used to handle the fact that PIT values are bounded on the interval [0, 1]. That is, any random steps across the boundaries are reflected back into the domain (Fig. 2). Mirror barriers are commonly used to describe stochastic processes in other areas of modeling [Karlin and Taylor (1981); see also Rose (1995) for applications in economics].

Let $p_{t_{\text{obs}}}$ be the PIT value of the most recent observation, and let $\Psi_n(p)$ be the probability density function of the verifying PIT value being p at n hours after t_{obs}. When $n = 0$, the PIT value is fully known and can therefore be described by

$$\Psi_0(p) = \delta(p - p_{t_{\text{obs}}}), \tag{6}$$

where δ is the Dirac delta function defined by

$$\delta(s) = \begin{cases} +\infty & s = 0 \\ 0, & s \neq 0 \end{cases} \quad \text{and} \tag{7}$$

$$\int_{-\infty}^{\infty} \delta(s)\, ds = 1. \tag{8}$$

Let $S(p, q)$ represent the probability density of arriving at a PIT value of p, given that the previous PIT value was q. Since our stochastic model for PIT values is a first-order Markov model, the probability of a certain PIT at time n can be found from all transitions to that PIT from time $n - 1$. The probability density after a transition can therefore be determined by the following recursive equation:

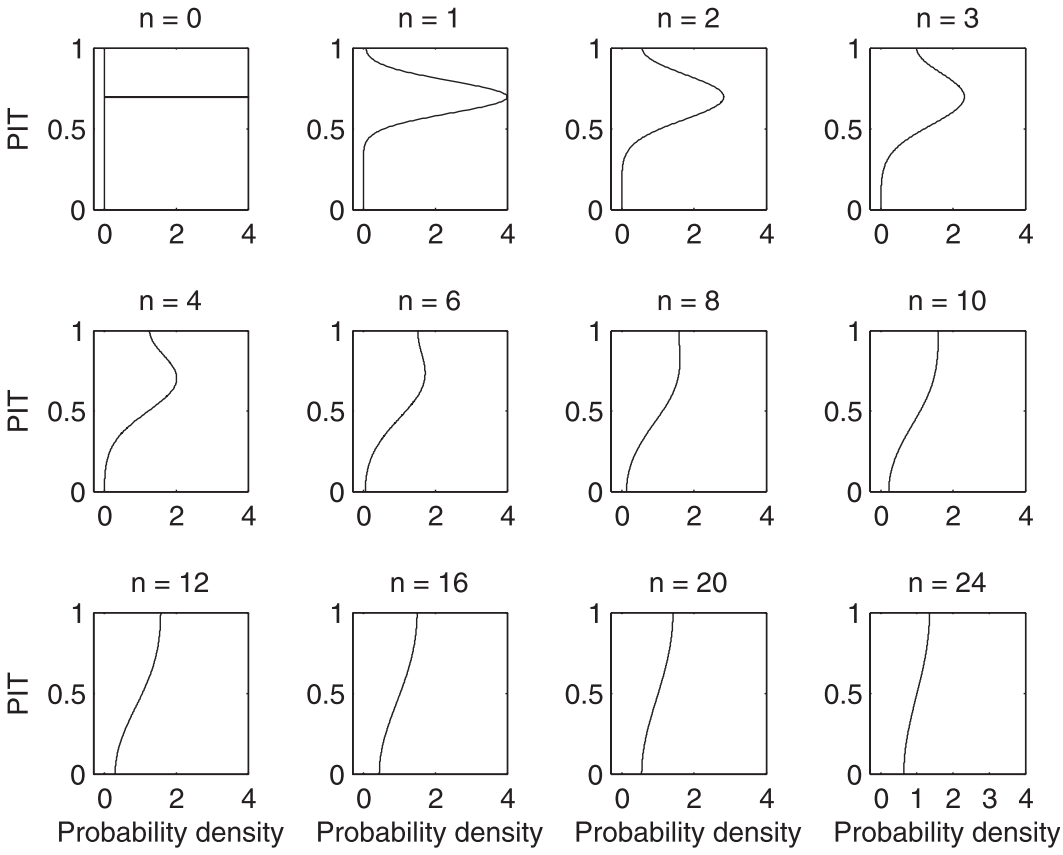

FIG. 3. An example sequence of PDFs of PIT values for different numbers of hours (n) after an observation has been made. In this case at $n = 0$, the PIT value is fully known to be 0.7.

$$\Psi_n(p) = \int_0^1 S(p,q)\Psi_{n-1}(q)\,dq. \tag{9}$$

b. Determining the transition function

We assume that the step length from one PIT to the next is Gaussian distributed with mean 0 and variance σ^2. That is, the transition function S can be constructed as follows:

$$S(p,q) = \phi(p;q;\sigma^2) + \phi(-p;q;\sigma^2) + \phi(2-p;q;\sigma^2) + \cdots \tag{10}$$

$$= \sum_{i=-\infty}^{+\infty} [\phi(p+2i;q;\sigma^2) + \phi(-p+2i;q;\sigma^2)], \tag{11}$$

where $\phi(x;\mu;\sigma^2)$ is a Gaussian PDF with mean μ and variance σ^2. The first term in Eq. (10) comes from steps within the domain, the second comes from steps reflected across 0, and the third term comes from steps reflected across 1. Equation (11) includes all possible steps, including steps that cross both boundaries one or more times.

A transition function that combines n number of steps can also be constructed and is denoted by S_n. The variance of multiple steps (under the assumed model) increases linearly with time, and S_n can therefore be computed by

$$S_n(p,q) = \sum_{i=-\infty}^{+\infty} [\phi(p+2i;q;n\sigma^2) + \phi(-p+2i;q;n\sigma^2)]. \tag{12}$$

Since σ is small in our study (around 0.15), and we use values of n no larger than 24, we restrict the summation to $i \in [-10, 10]$. A wider range for i may be required for large σ and n values.

Constructing S_n allows us to simplify Eq. (9) to the following:

$$\Psi_n(p) = \int_0^1 S_n(p,q)\Psi_0(q)\,dq \tag{13}$$

$$= S_n(p, p_{t_{\text{obs}}}), \tag{14}$$

where again $p_{t_{\text{obs}}}$ is the verifying PIT value at time t_{obs}. This simplification avoids the need to recursively compute Ψ_n [as in Eq. (9)]. Note that for forecast variables that require a non-Gaussian transition function, it is possible that Eq. (12) cannot be constructed analytically in which case the above simplification may not be possible.

Figure 3 shows an example sequence of $\Psi_n(p)$ for various values of n. The PIT value distribution clearly widens as time goes on, indicative of the disappearing effects of the last observed PIT value.

c. Parameter estimation

To create the updated forecasts, an estimate of σ^2 is needed by Eq. (12). The variance of the step sizes of past PIT values (σ_0^2) can be used:

$$\sigma_0^2 = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (p_{t+1} - p_t)^2, \tag{15}$$

where $\mathcal{T}$ represents a set of time points from the past forecast cycles that compose the training period and where $|\mathcal{T}|$ is the size of this training set. In general, σ_0^2 will underestimate σ^2 since some steps will appear to be short steps when in fact they are longer steps that have reflected across a boundary.

For a given σ, the expected value of σ_0 can be computed by the integral over all possible PIT transitions from p to q:

$$\sigma_0^2 = \sum_{i=-\infty}^{\infty} \int_0^1 \int_0^1 [\phi(p + 2i; q; \sigma)(p - q)^2 + \phi(-p + 2i; q; \sigma)(p - q)^2]\, dp\, dq. \tag{16}$$

Solving this equation for σ [as required by Eq. (12)] was not possible analytically. We found that the following is a good approximation for σ in terms of σ_0:

$$\sigma \approx \tan(3.5\sigma_0)/3.5, \tag{17}$$

where the input to the tangent function is in radians. This approximation has errors of less than 3.4% for σ_0 values up to 0.3 (Fig. 4).

A summary of the process involved with updating a probabilistic forecast goes as follows: the variance of past PIT transition distances (σ_0) is computed by Eq. (15), which is used to approximate σ in Eq. (17); σ is then used in Eq. (12) to compute the transition function (S_n); and the transition function, combined with the latest available verifying PIT value, are used to calculate the PIT distribution (Ψ_n) by Eq. (14), which is used to update the original probabilistic forecast through Eq. (5).

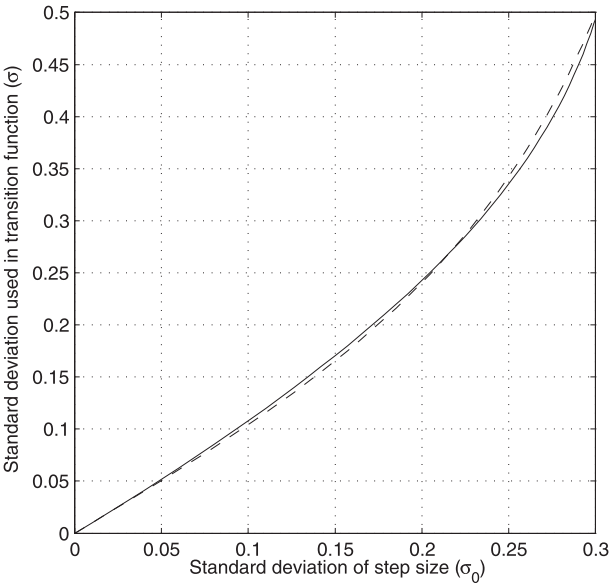

FIG. 4. Standard deviation of PIT step sizes used in the transition function as a function of the measured standard deviation of step sizes of past PIT values (solid line) and the approximation $\sigma = \tan(3.5\sigma_0)/3.5$ (dashed line).

3. Operational test case

a. Model data and configuration

Hourly surface temperature forecasts from the Mesoscale Compressible Community (MC2; Benoit et al. 1997) model, the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (Penn State–NCAR) Mesoscale Model (MM5; Grell et al. 1994), and the Weather Research and Forecasting Model (WRF; Skamarock et al. 2005) were used for the case study period: 0000 UTC 1 September 2005–2300 UTC 1 February 2008. Two runs for the WRF model were used: one using Global Forecast System (GFS) initialization (WRFG) and the other using North American Mesoscale Modeling (NAM) model initialization (WRFN), while MC2 and MM5 both used NAM initialization. The MC2 and MM5 runs had outer domains with 108-km grid spacing, and inner 36-, 12-, and 4-km nested domains. The WRF runs were similar, but did not contain the 4-km nested domain. These domains composed our 14-member ensemble.

The models were initialized once per day at 0000 UTC, and hourly forecast output to 60 h was available. Probabilistic forecasts were generated for the same time period.

The model runs and probabilistic forecasts were generally completed by 0600 UTC, after which we used the latest observation to update the probabilistic forecasts valid for the subsequent 24 h. The update process

was repeated each hour as a new observation became available. This was done until 0600 UTC the next day, when the probabilistic forecasts from the next forecast cycle were used. This means that for each forecast cycle twenty-four 24-h updated forecasts were produced, yielding 576 forecasts per day.

We tested the method on temperature probabilistic forecasts and observations for the following five airport stations in British Columbia, Canada: Vancouver International Airport station (CYVR), Abbotsford International Airport (CYXX), Victoria International Airport (CYYJ), Kamloops Airport (CYKA), and Kelowna Airport (CYLW). This group of stations provided a geographically diverse sample from within our smallest model domain.

b. Original probabilistic forecasts

We used the method of moments to produce the original probabilistic forecast from the forecast ensemble. The PDF using this method is computed by

$$f_t(x) = \phi(x; \xi_t + \mu; s^2), \tag{18}$$

where again ϕ is a Gaussian PDF, x is a temperature value, ξ_t is the ensemble mean at time t, μ is a bias-correction term for the center of the distribution, and s^2 is the variance of the distribution.

The last two parameters are determined by the forecast errors during the training period $\mathcal{T}$:

$$\mu = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} x_i - \xi_i \quad \text{and} \tag{19}$$

$$s^2 = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} (x_i - \mu - \xi_i)^2. \tag{20}$$

Note that the spread in this case is independent of the ensemble spread.

The parameters μ and s were computed separately for each station and separately for each of the 24 forecast hours. They were computed from a 40-day sliding window that ended the day before the forecast was initialized. A training period of 40 days is a compromise between the need to use statistics that adapt quickly to seasonal changes and the requirement to have enough data to robustly estimate the parameters. Similar training lengths have been used to produce probabilistic forecasts using Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2007).

The spread parameter σ_0 (and consequently σ) was also computed separately for each station using a 40-day sliding window; however, all 24 forecast offsets for a given station were pooled together to give a more robust estimate.

4. Analysis

a. Ignorance score

We use the logarithmic score of Good (1952), which has gained popularity over the last decade and has been referred to as the ''ignorance'' score owing to its ties with information theory (Roulston and Smith 2002). It is defined as follows:

$$\mathrm{IGN}(f) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} -\log_2[f_t(x_t)]. \tag{21}$$

IGN rewards forecasts that place high confidence in the value where the observation falls. Low ignorance scores are desired.

The total ignorance scores of the original probabilistic forecasts were computed by averaging ignorance scores over all forecast cycles, and forecast hours, but separately for each station and each value of n in order to see how far into the future a recent observation can improve the ignorance score.

Figure 5a shows the improvement in the ignorance score provided by the updated probabilistic forecast as a function of distance from the most recent observation. The updated forecasts at 0 h after an observation has been made has an ignorance score of $-\infty$ since the true state is fully known. However, this update forecast is of no value since it is only available after the observation has been made. As the time since the most recent observation increases, the improvement in the ignorance score reduces down toward 0.

b. Continuous ranked probability score

We also computed the continuous ranked probability score (CRPS) to further evaluate the quality of the probabilistic forecasts. It is defined as

$$\mathrm{CRPS}(F) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \int_{-\infty}^{+\infty} [F_t(x) - H(x - x_t)]^2 \, dx, \tag{22}$$

where $H(s)$ is the Heaviside function defined by

$$H(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases}. \tag{23}$$

Low values of CRPS are preferred.

Figure 5b shows the percentage improvement due to the updated forecast relative to the original raw forecast. This is defined as

$$\%\mathrm{improvement} = \frac{\mathrm{CRPS}(F_{\mathrm{raw}}) - \mathrm{CRPS}(F_{\mathrm{updated}})}{\mathrm{CRPS}(F_{\mathrm{raw}})} \times 100\%. \tag{24}$$

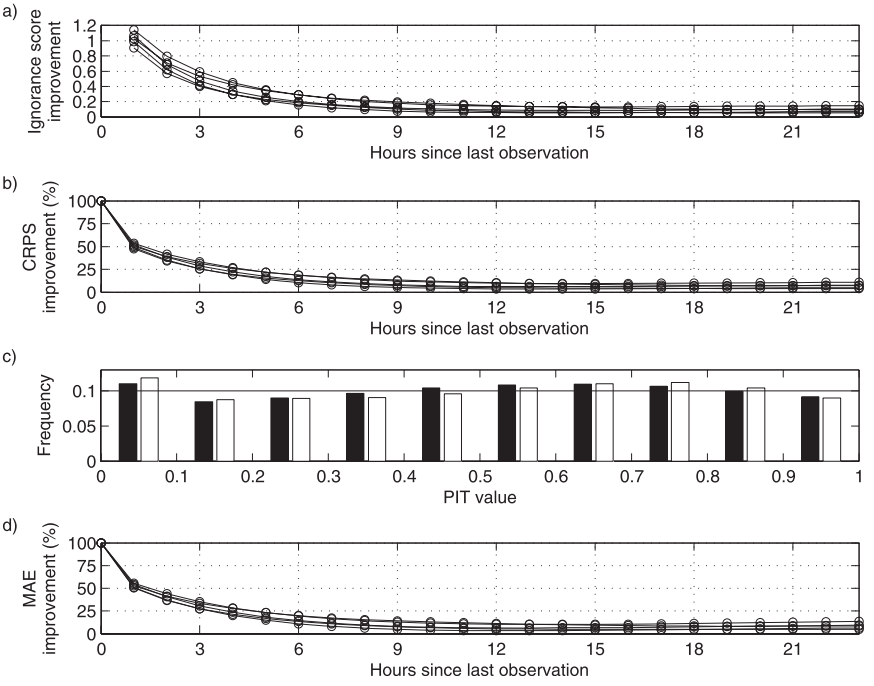

FIG. 5. Verification statistics for the probabilistic forecasts used in the study. (a) Reduction (improvement) of the ignorance score by the updated probabilistic forecast relative to the original probabilistic forecast. Each of the five lines represents the score for a different station. (b) Percentage improvement in the CRPS by the updated probabilistic forecast. (c) PIT histogram of the updated forecasts (black bars) and the original forecasts (white bars), indicating the reliability of the forecasts. (d) Percentage improvement in mean absolute error of the median of the updated probability distributions relative to the median of the original distribution.

Results for CRPS show a similar pattern as for the ignorance score, with the update method providing less improvement as the time since the most recent observation increases. The average CRPS of the five stations was 1.50°C and the update method brought the values down to 1.06° and 1.27°C at 3 and 6 h, respectively.

c. Reliability

A probabilistic forecast is reliable (or calibrated) when the PIT values are uniformly distributed between 0 and 1 (Gneiting et al. 2007). This can be diagnosed by a simple histogram of verifying PIT values, as reliable forecasts will give a flat histogram.

Figure 5c shows the histogram of PIT values from all forecast hours, forecast cycles, stations, and values of n. The update method does not appear to degrade or improve the reliability of the original forecasts in any significant way.

d. Mean absolute error

A probabilistic forecast can also provide a best deterministic estimate, by using the median of the probability distribution (as shown by the 50% lines in Figs. 1a and 1b). We used the mean absolute error (MAE) as a verification measure of this deterministic forecast:

$$\mathrm{MAE}(f) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} |x_t - F_t^{-1}(0.5)|, \qquad (25)$$

where F_t^{-1} is the inverse of F_t giving the temperature value corresponding to a nominal proportion of 0.5.

The MAE of the deterministic forecast (Fig. 5d) showed a similar pattern to the ignorance score and CRPS, with the update method improving the MAEs from 2.07°C down to 1.42°C and 1.73°C at 3 and 6 h, respectively. Improvements in MAE suggest that the update method improves the central tendency of the probabilistic forecasts.

5. Conclusions

We have presented a method to update probabilistic forecasts of continuous variables based on recent observations, which should prove useful for a variety of nowcasting purposes. An alternative to this is to use data assimilation after new observations are available in

order to create new initializations for the ensemble, followed by a complete rerun of the ensemble. This is considerably more expensive from a computational point of view, and may be infeasible for many operational systems.

The method improves the ignorance score and CRPS of the probabilistic forecasts, and improves the MAE of the median of the distribution significantly for forecasts up to 6 h after a recent observation, while not affecting reliability negatively.

Future work includes investigating the benefits of using a higher-order Markov model for modeling PIT transitions. In addition to accounting for the hour-by-hour correlation of PIT values, a higher-order Markov model can also incorporate any diurnal correlation of PIT values that may exist, thereby allowing for the potential to improve forecasts for 24 h after a recent observation.

Acknowledgments. This research was made possible by funding from the Canadian Natural Science and Engineering Research Council, the Canadian Foundation for Climate and Atmospheric Science, and the BC Hydro and Power Authority. We also thank Kristian Soltesz and two reviewers for providing helpful insight.

REFERENCES

AMS, 2008: Enhancing weather information with probability forecasts. *Bull. Amer. Meteor. Soc.,* **89,** 1049–1053.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic precipitation forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Benoit, R., M. Desgagne, P. Pellerin, S. Pellerin, Y. Chartier, and S. Desjardins, 1997: The Canadian MC2: A semi-Lagrangian, semi-implicit wideband atmospheric model suited for finescale process studies and simulation. *Mon. Wea. Rev.,* **125,** 2382–2415.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.,* **132,** 338–347.

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.,* in press.

Glahn, H., and D. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **14B,** 107–114.

Grell, G. J., J. Dudhia, and D. R. Stauffer, 1994: A description of the fifth generation Penn State/NCAR Mesoscale Model (MM5). NCAR Tech. Rep. TN-398+STR, 122 pp.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

Homleid, M., 1995: Diurnal correction of short-term surface temperature forecasts using the Kalman filter. *Wea. Forecasting,* **10,** 689–707.

Jewson, S., A. Brix, and C. Ziehmann, 2005: *Weather Derivative Valuation*. Cambridge University Press, 373 pp.

Karlin, S., and H. Taylor, 1981: *A Second Course in Stochastic Processes*. Academic Press, 582 pp.

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.

Rose, C., 1995: A statistical identity linking folded and censored distributions. *J. Econ. Dyn. Control,* **19,** 1391–1403.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Rep. TN-468+STR, 88 pp.

Sloughter, J. M., A. E. Raftery, and T. Gneiting, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.