Probabilistic Forecasting of Snowfall Amounts Using a Hybrid between a Parametric and an Analog Approach

MICHAEL SCHEUERER

Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, and Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

THOMAS M. HAMILL

Physical Sciences Division, NOAA/Earth System Research Laboratory, Boulder, Colorado

(Manuscript received 2 August 2018, in final form 6 December 2018)

ABSTRACT

Forecast uncertainty associated with the prediction of snowfall amounts is a complex superposition of the uncertainty about precipitation amounts and the uncertainty about weather variables like temperature that influence the snow-forming process. In situations with heavy precipitation, parametric, regression-based postprocessing approaches often perform very well since they can extrapolate relations between forecast and observed precipitation amounts established with data from more common events. The complexity of the relation between temperature and snowfall amounts, on the other hand, makes nonparametric techniques like the analog method an attractive choice. In this article we show how these two different methodologies can be combined in a way that leverages the respective advantages. Predictive distributions of precipitation amounts are obtained using a heteroscedastic regression approach based on censored, shifted gamma distributions, and quantile forecasts derived from them are used together with ensemble forecasts of temperature to find analog dates where both quantities were similar. The observed snowfall amounts on these dates are then used to compose an ensemble that represents the uncertainty about future snowfall. We demonstrate this approach with reforecast data from the Global Ensemble Forecast System (GEFS) and snowfall analyses from the National Operational Hydrologic Remote Sensing Center (NOHRSC) over an area within the northeastern United States and an area within the U.S. mountain states.

1. Introduction

Snow forecasts are not only of interest for recreational activities like skiing, they are also vital for planning and decision-making in various sectors of the economy such as air and ground transportation, agriculture, construction, and commerce. Especially in regions with complex terrain like the western United States, convection-permitting high-resolution ($\leq$4 km) models are needed to adequately represent the spatial differences in precipitation between mountains and valleys (Gowan et al. 2018). Current operational high-resolution guidance like the North American Mesoscale Forecast System 3-km continental U.S. nest (Rogers et al. 2017) is only available for up to 60 h ahead, and so it is useful to explore statistical postprocessing algorithms that can correct coarser medium-range (i.e., several days ahead) snowfall forecasts to reflect expected finer-scale variability and represent the associated forecast uncertainty.

The uncertainty around medium-range forecasts of precipitation amounts like those from NOAA's Global Ensemble Forecast System (GEFS; Zhou et al. 2017) can be substantial (Scheuerer and Hamill 2015, their Fig. 6). For snowfall amounts, additional uncertainty about the precipitation type (rain, snow, etc.) and the snow-to-liquid ratio (SLR) if the precipitation type is snow must be taken into account. This challenge is illustrated in Fig. 1 where both analyzed precipitation (i.e., liquid water) amounts and analyzed snowfall over Utah, western Colorado, and southern Wyoming are shown for a 24-h period ending at 1200 UTC 18 April 2015.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-18-0273.s1.

Corresponding author: Michael Scheuerer, michael.scheuerer@noaa.gov

DOI: 10.1175/MWR-D-18-0273.1

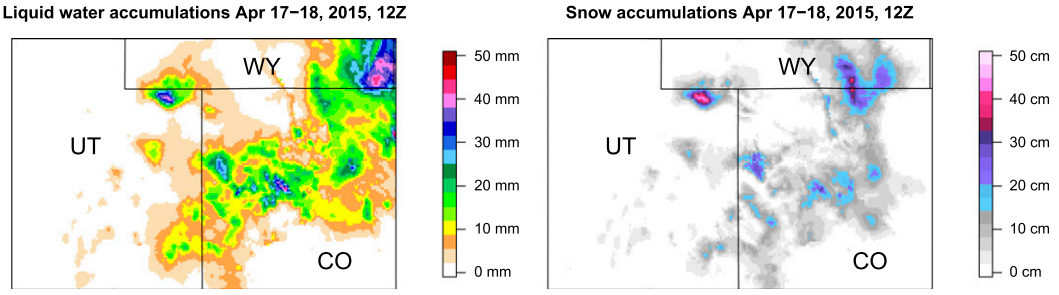

FIG. 1. Analyzed liquid water accumulations (Stage-IV precipitation analyses) and analyzed snowfall amounts (NOHRSC snowfall analyses) over Utah, western Colorado, and southern Wyoming for the 24-h period ending at 1200 UTC 18 Apr 2015.

Over the mountainous regions, temperatures were cold enough for precipitation to fall as snow. At the lower elevations in southeastern Wyoming and north-central Colorado (near the right margin of both plots) heavy precipitation but little or no snow accumulation was reported. Forecast uncertainty about the temperature in these areas translates into uncertainty about snow accumulations that could have been quite large if temperatures were low enough for the heavy precipitation to fall as snow instead of rain. The coarse resolution of topography in global forecast systems does not permit an adequate representation of the spatial variability of temperature in complex terrain (Dabernig et al. 2017, their Fig. 4), and we may therefore expect that the explicitly modeled snow accumulations produced by modern operational systems (e.g., Zhao and Carr 1997) at the medium range suffer from even stronger local biases than forecasts of precipitation amounts. These systematic biases could be addressed by statistical postprocessing of explicitly modeled snow accumulations (if available) directly, but by doing so the different sources of forecast uncertainty about temperature and precipitation amounts would be lumped together. Moreover, it would be difficult to guarantee consistency between forecasts of precipitation and snow accumulations, and we therefore prefer to address biases in precipitation and temperature forecasts separately.

The overall uncertainty about snow accumulation is a combination of the uncertainty about precipitation accumulation, precipitation type, and SLR. One of the challenges is that these three quantities are not independent of each other. Figure 2a) shows a scatterplot of SLR versus snow water equivalent (SWE) at the analysis grid point closest to Alta ski resort. SWE and SLR were derived based on Stage-IV precipitation analyses and NOHRSC snowfall analyses (see section 2 for details about these datasets), only days with at least 2.5 mm of accumulated precipitation and at least 5 cm of accumulated snow were considered. Both Fig. 2a and a similar plot based on observed precipitation and snow data at Collins Snow Study Plot (CLN) at Alta ski resort (Alcott and Steenburgh 2010, their Fig. 5e) show that SLR tends to decrease as the amount of SWE increases. This tendency can be observed at most of the analysis

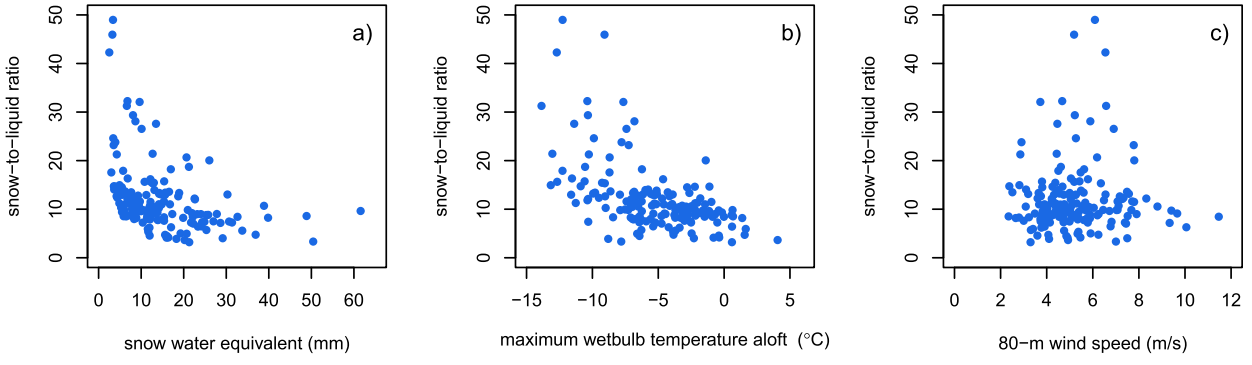

FIG. 2. Scatterplots of snow water equivalent (Stage-IV precipitation analyses), maximum wet-bulb temperatures aloft, and 80-m wind speeds (both based on 12–36-h GEFS forecasts) against SLR at the analysis grid point associated with Alta ski resort in Utah. SLR was calculated based on Stage-IV precipitation analyses and NOHRSC snowfall analyses, only days with at least 2.5 mm of accumulated precipitation and at least 5 cm of accumulated snow were considered.

grid points with a sufficient number of snow days and was also noted by Ware et al. (2006), who studied the relation between SLR and various predictors including SWE at 28 stations across the United States. Part of the uncertainty about snowfall predictions therefore cancels out (e.g., if SWE amounts are higher than expected, this also entails higher snowfall amounts, but the increase is less than proportionally due to the simultaneous decrease in SLR), and this should be taken into account when constructing a probabilistic snow forecast.

The most important predictor for determining the precipitation type and SLR is the temperature in the snow-forming layer (Ware et al. 2006; Alcott and Steenburgh 2010). Following up an earlier study by Roebber et al. (2003), Ware et al. (2006) consider a low–midlevel temperature factor derived using a principal component analysis of vertical temperature profiles. Alcott and Steenburgh (2010) studied the correlation between SLR at Alta ski resort (where the surface pressure is around 690 hPa) and the temperature at various pressure levels, and they found a strong and near-constant correlation across all available levels from 850 to 400 hPa. The specific choice of the temperature predictor considered in our study, the maximum wet-bulb temperature between 609.6 m (2000 ft) above ground level and 400 hPa, is compatible with these findings and was made because this quantity (''maximum wet-bulb temperature aloft'') is one of the weather elements included in NOAA's National Blend of Models (NBM), a nationally consistent and skillful suite of calibrated forecast guidance in which the algorithm proposed in this paper may be included in the future.

Other predictors that can be informative for SLR are wind speed and relative humidity (Ware et al. 2006; Alcott and Steenburgh 2010). Alcott and Steenburgh (2010) note that the correlation between wind speed and SLR peaks at 650 hPa (600 hPa on the subset of high SWE events), and we therefore consider 80-m wind speeds as a predictor in our study as these are the closest match to the findings by Alcott and Steenburgh (2010) among the available NBM weather elements. For relative humidity, our attempt to be compatible with available NBM weather elements leaves us with 2-m relative humidity as the only option. Figure 2b shows that the maximum wet-bulb temperature aloft predictor (here: interpolated 12–36-h forecasts by the GEFS as detailed below) is related to SLR in a rather nonlinear and heteroscedastic way. For the ''80-m wind speed'' predictor Fig. 2c suggests that there is barely any evident relationship with SLR. If probabilistic forecasts are to be generated where not just the mean but also the uncertainty about SLR needs to be explained by these predictors, it is clear that constructing a parametric statistical model for that purpose is rather challenging; this suggests that a nonparametric approach like an analog method (Sievers et al. 2000; Hamill and Whitaker 2006; Delle Monache et al. 2013) is a preferable choice. On the other hand, using a parametric approach to postprocess NWP forecasts of precipitation amounts has advantages over an analog method since it represents the distribution better (an analog method introduces sampling variability) and it can extrapolate relations between forecast and observed precipitation amounts established with data from common events to situations with more extreme precipitation (Scheuerer and Hamill 2015). We propose a new approach that combines an analog method with the parametric postprocessing technique proposed by Scheuerer and Hamill (2015) in a way that leverages the respective advantages.

A recent paper by Stauffer et al. (2018) also discusses probabilistic prediction of snowfall amounts and uses a parametric postprocessing approach for both precipitation and temperature. In contrast to the approach presented here, their paper addresses the additional challenge of disaggregating predicted snow accumulations to an hourly temporal scale, and it derives snowfall amounts based on a fixed temperature threshold to determine the precipitation type and a fixed value for SLR. The Stauffer et al. (2018) approach is not predicated on the availability of high-quality snowfall observations, it only requires observations of temperature and precipitation amount and is therefore more widely applicable. If dependable snowfall observations or analyses are available, however, the methods presented below allow one to use that information for addressing the complex relationship between temperature, precipitation type, and SLR, and they can potentially yield more accurate forecasts and a better representation of forecast uncertainty than is possible with a fixed temperature threshold for rain versus snow and a fixed value for SLR.

In section 2 we describe the forecast and analysis data used in this study. Section 3 briefly reviews the analog method and the censored shifted gamma distribution approach proposed by Scheuerer and Hamill (2015), and explains how those two techniques can be combined in order to generate an ensemble of snowfall forecasts based on forecasts of precipitation, temperature, etc. by the Global Ensemble Forecast System (GEFS). The resulting probabilistic forecasts of snowfall accumulations are evaluated in section 4, and are compared to forecasts generated with the standard analog method. We finally discuss some of the limitations of our method in section 5. The analyses in this study have been performed using the statistical software R (R Core Team 2017).

2. Data used in this study and data preprocessing

The postprocessing methodology discussed here is applied to GEFS ensemble forecasts during the period from October 2013 to May 2017. Forecast data were obtained from the second-generation GEFS reforecast dataset (Hamill et al. 2013), which consists of 11 ensemble member forecasts, initialized at 0000 UTC every day. The following weather variables were extracted from this dataset:

- 6-h precipitation accumulations on a ~½° Gaussian grid for forecast lead times up to 132 h;
- *u*/*v* components of wind at 80 m on a ~½° Gaussian grid for forecast lead times up to 132 h in 6-h increments;
- ensemble mean 2-m temperature forecasts on a 1° Gaussian grid for forecast lead times up to 132 h in 6-h increments;
- ensemble mean surface pressure forecasts on a 1° Gaussian grid for forecast lead times up to 132 h in 6-h increments;
- ensemble mean 2-m specific humidity forecasts on a 1° Gaussian grid for forecast lead times up to 132 h in 6-h increments;
- ensemble mean temperature forecasts on a 1° Gaussian grid at pressure levels 1000, 925, 850, 700, and 500 hPa for forecast lead times up to 132 h in 6-h increments;
- ensemble mean geopotential height forecasts on a 1° Gaussian grid at pressure levels 1000, 925, 850, 700, and 500 hPa for forecast lead times up to 132 h in 6-h increments; and
- ensemble mean specific humidity forecasts on a 1° Gaussian grid at pressure levels 1000, 925, 850, 700, and 500 hPa for forecast lead times up to 132 h in 6-h increments.

The *u*/*v* wind components were converted to 80-m wind speed forecasts, averaged over the 11 ensemble members, and bilinearly interpolated to the ~0.04° snowfall analysis grids (see below). Surface temperature, pressure, and specific humidity forecasts were used to calculate relative humidity forecasts and bilinearly interpolated to the ~0.04° grid. Temperature and specific humidity forecasts at the different pressure levels were used to calculate wet-bulb temperatures. These wet-bulb temperatures and the geopotential heights were then bilinearly interpolated to the ~½° Gaussian grid, and at this resolution the maximum wet-bulb temperature aloft predictor was calculated. This quantity is one of the weather elements included in NOAA's National Blend of Models (NBM) and defined as the maximum temperature between 609.6 m (2000 ft) above ground level (AGL) and the 400-hPa level. The wet-bulb temperature at 609.6 m AGL was determined by vertical linear interpolation of both wet-bulb temperature and geopotential height forecasts. The model grid surface elevation was used as a baseline with respect to which the height AGL is calculated, and the interpolated temperature value corresponding to the interpolated geopotential height of 609.6 m AGL was selected. The maximum wet-bulb temperature aloft was then estimated as the maximum over this value and the wet-bulb temperature at all pressure levels up to 500 hPa with a geopotential height larger than 609.6 m AGL. Finally, this maximum wet-bulb temperature aloft predictor was bilinearly interpolated to the ~0.04° snowfall analysis grid.

Gridded snowfall data used in our study comes from National Operational Hydrologic Remote Sensing Center (NOHRSC) snowfall analyses on a ~0.04° Gaussian grid over the conterminous United States. For the time period studied here (2013–17) these analyses are currently only available for 24-h accumulation periods beginning and ending at 1200 UTC. They use Stage IV precipitation analyses (Lin 2011) to establish the total precipitation amount associated with the first-guess snowfall. Snow-to-precipitation ratios are derived from either NOAA's High-Resolution Rapid Refresh (HRRR) model if HRRR has nonzero precipitation consistent with the Stage IV precipitation analyses. Otherwise, a climatological SLR and a temperature-based decision rule for snow versus rain is used. This first guess is then updated with quality-controlled snowfall observations in two assimilation passes: the first pass mainly accounts for bias and SLR differences, the second pass interpolates the actual difference between first-pass snowfall and observed snowfall. For further details about the NOHRSC, version 2, snowfall analyses see Clark (2017). For the hybrid postprocessing method proposed below, we further need analyses of 6- and 24-h precipitation amounts, and the Stage IV dataset is a natural choice since the first guess of the NOHRSC snowfall analyses is based on these data.

Since snowfall analyses were only available for 24-h accumulation periods, the predictors have to be aggregated to this time scale. For precipitation this is done by simply summing up the amounts predicted for the four respective 6-h periods. For temperature and wind speed, simple averaging may not be appropriate since both quantities may vary over the course of 24-h period. Following a suggestion by T. Alcott (2017, personal communication), we weighed the four 6-h periods proportional to the sum of the climatological average precipitation and the actual precipitation amount in each of these periods. The resulting 24-h aggregation of temperature and wind speed emphasizes the 6-h period(s) where most precipitation occurs. During the training periods, we use 6-h Stage IV precipitation accumulations

as the "actual" precipitation amount that determines the weights; during the verification periods our best estimate of the actual precipitation amount is based on the GEFS ensemble forecasts. These come with substantial uncertainty, especially at longer lead times, and so we use the ensemble mean of the augmented (see section 3b) GEFS ensemble to smooth out unwarranted spatial detail in the weights. Timing errors in the precipitation forecasts can still result in suboptimal weights; this is why we base the weights partly on climatological average precipitation. If the forecasts were known to suffer from systematic timing errors, it might be preferable to use forecast precipitation amounts to calculate the weights during both training and verification period. If timing errors are nonsystematic, however, this alternative weighting strategy would introduce an error at two places, and we therefore chose to calculate the weights during the training period based on analyzed precipitation amounts so that at least temperatures during the training period are weighed optimally.

Both forecast and analysis datasets are composed for two areas within the United States: one of them ("mountain region") is the area shown in Fig. 1 covering Utah, western Colorado, and southern Wyoming, the other area ("northeast region") covers several states in the northeastern United States near the Atlantic coast (see Fig. 5).

3. Statistical postprocessing methodology

a. Analog method

We use an analog method for statistical postprocessing (Sievers et al. 2000; Hamill and Whitaker 2006; Delle Monache et al. 2013) as a reference method and as one of two components of the hybrid approach proposed in section 3c. The main idea behind this technique is simple, yet very effective:

- choose a set of past dates where the values of the predictors were similar to those associated with today's forecast, and
- form an analog forecast ensemble from the values of the predictand on these dates.

In the present setup the predictand is 24-h snow accumulation and the predictors are the GEFS ensemble mean forecasts of 24-h precipitation accumulations (Pa) and the 24-h weighted average (see section 2) of 6-h maximum wet-bulb temperatures aloft (Ta). Additional predictors like wind speed or relative humidity can be added in a straightforward way, their utility is studied in section 4e. The set of analog dates is chosen separately for each grid point and lead time (using only forecasts from that specific grid point), but since the interpolated coarse-scale predictors are relatively smooth in space we can expect that a similar set of analog events will be chosen for nearby points. Denoting the predictor values associated with today's forecast with a superscript "t" and those associated with the forecast values at some historic date d with a superscript "d," we quantify the dissimilarity Δ_d between the respective forecasts via

$$\Delta_d = w_{\mathrm{Pa}}\left|\sqrt{\mathrm{Pa}^d} - \sqrt{\mathrm{Pa}^t}\right| + w_{\mathrm{Ta}}|\mathrm{Ta}^d - \mathrm{Ta}^t|. \qquad (1)$$

Although it is common practice to use standardized predictors for defining a dissimilarity measure, we chose not to do that in the present context since we felt that it is more important to account for the very different nature of the distributions of the different predictors. Specifically, to address the high skewness in the distributions of precipitation accumulations, we apply a square root transformation before calculating absolute differences. The coefficients w_{Pa} and w_{Ta} allow one to emphasize predictors that are thought to have the strongest link with snowfall amounts, and they can also compensate for differences in the weighting of predictors that result from their different scales that were not accounted for by standardization. Without loss of generality we can fix $w_{\mathrm{Ta}} = 1$, and we study the effect of different choices of w_{Pa} in section 4a.

Another tuning parameter that affects the performance of an analog method is the size of the analog ensemble (i.e., the number n_{ana} of similar past dates chosen by an analog algorithm). While larger ensembles reduce sampling variability and thus yield a better representation of the forecast distribution, additional members are increasingly dissimilar from today's forecast and can therefore be detrimental to the accuracy of the analog ensemble. Hamill et al. (2006) applied an analog method to postprocess precipitation reforecasts from a previous version of NCEP's Global Forecast System and show that the optimal ensemble size depends on various factors including training sample size (see their Fig. 7). In section 4a we therefore test different analog ensemble sizes and study their role in the performance of the analog method for snowfall amounts.

b. Nonhomogeneous regression based on censored shifted gamma distributions

One of the challenges with an analog method is that for unusual values of the predictors (e.g., extreme precipitation amounts) it becomes very difficult to find good analog dates. When several predictors are considered simultaneously, and analog dates with sufficiently similar forecast values for all predictors have to be found, this challenge becomes even harder. In these situations

parametric, regression-based postprocessing approaches have the advantage that they can extrapolate relations between forecast and observed values established with data from more common events. For precipitation amounts, Scheuerer and Hamill (2015) proposed an approach based on censored shifted gamma distributions that links the distribution parameters to statistics of the GEFS precipitation forecasts. In this study we use a slightly modified version of their method, described in section 3a of Scheuerer and Hamill (2018). The following is a brief review of the main ideas:

Ensemble augmentation: Instead of just using the ensemble forecasts at the nearest forecast grid point, we use all forecasts within a certain neighborhood of the location of interest. Forecast grid points within this neighborhood are weighted proportionally to their predictive skill as described in section 3.2 of Scheuerer et al. (2017b).

Ensemble statistics: The information in this augmented GEFS ensemble is summarized by three statistics: the ensemble probability of precipitation (POP), the ensemble mean (EM), and the ensemble mean absolute difference (MD, a measure of ensemble spread). Since the augmented ensemble contains forecasts from different grid points and thus different climatologies, the forecasts must be spatially homogenized before calculating the ensemble statistics. Here, we work with multiplicative forecast anomalies $\tilde{x}_i := x_i/\overline{x}_{\mathrm{cl}}$ derived from the 11 GEFS raw ensemble member forecasts $x_1, \ldots, x_{11}$ at each particular forecast grid point by dividing them by the climatological average $\overline{x}_{\mathrm{cl}}$ of the forecast precipitation amounts at this grid point.

Predictive distributions: Censored, shifted gamma distributions (CSGDs) are used to model forecast uncertainty. They are parameterized by the mean and standard deviation μ and σ of a gamma distribution, and by an additional shift parameter δ that shifts the distribution toward the negative values. The shifted distribution is then censored at zero, which allows one to model occurrence and intensity of precipitation simultaneously.

Climatological distributions: Since the ensemble statistics are based on multiplicative forecast anomalies, the local climatological information has to be added back in when the predictive distributions are defined. This is achieved by first fitting a climatological CSGD to the analyzed precipitation amounts at each grid point; the resulting parameters μ_{cl}, σ_{cl} and δ_{cl} are included in the regression equations that determine the predictive CSGD parameters.

Regression equations: The GEFS ensemble statistics are finally linked to the predictive CSGD parameters via

$$\mu = \frac{\mu_{\mathrm{cl}}}{a_1}\log(1 + \{[\exp(a_1) - 1](a_2 + a_3\mathrm{POP} + a_4\mathrm{EM})\}),$$

$$\sigma = \sigma_{\mathrm{cl}}\left(b_1\sqrt{\frac{\mu}{\mu_{\mathrm{cl}}}} + b_2\mathrm{MD}\right). \tag{2}$$

The shift parameter is kept fixed at $\delta = \delta_{\mathrm{cl}}$.

A separate set of model parameters is fitted for each analysis grid point, each forecast lead time, and each month, using data from ±45 days around the 15th of the respective month during the years set aside for training. This way, a sufficient amount of training data is available to permit a stable estimation of the model parameters, while allowing the fitted model to vary in space and over the course of the year in order to address spatially and seasonally varying biases. Scheuerer and Hamill (2015) demonstrate that this regression-based approach yields sharper forecast distributions than an analog method for precipitation amounts, and is more likely to preserve a strong forecast signal in the raw GEFS ensemble forecasts while still being statistically reliable.

c. *Combining the CSGD and an analog method*

We now show how the two approaches presented above—analog method and CSGD-based nonhomogeneous regression method—can be combined in order to leverage the respective strengths. The general idea can be summarized as follows:

- the dissimilarity measure Δ_d in (1) used to selecting the analog dates depends on the Ta predictor in the same way as before (i.e., dissimilarity is defined by comparing today's and historic GEFS ensemble mean forecasts);
- the Pa predictor, on the contrary, is first postprocessed as described in section 3b, and dissimilarity is defined by comparing the values of a *sample from the resulting predictive distribution* for Pa with *analyzed* Pa at the historic dates; and
- analog dates are found separately for the different Pa sample values, and a predictive snowfall ensemble is composed of the snowfall amounts on these dates.

For a more formal description, denote by $x_\star$ and $y_\star$ forecast and analyzed quantities, respectively, where $\star$ is either "Snow," "Pa," or "Ta." Further denote by $p(y_{\mathrm{Snow}}|x_{\mathrm{Pa}}, x_{\mathrm{Ta}})$ the conditional probability distribution function (PDF) of y_{Snow} given x_{Pa} and x_{Ta}. Under the following assumptions:

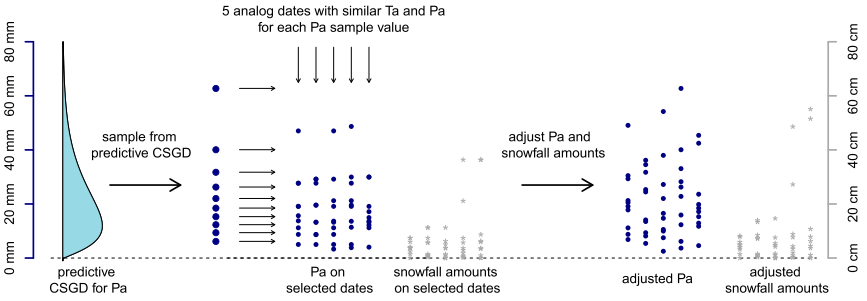

FIG. 3. Schematic illustration of how the parametric CSGD method is combined with an analog method in order to generate a forecast ensemble of snowfall amounts based on the predictive distribution for precipitation amounts and GEFS ensemble mean forecasts of maximum wet-bulb temperature aloft.

- the predictor x_{Ta} is negligible for forecasting y_{Pa};
- given x_{Ta} and y_{Pa}, the predictor x_{Pa} has no additional information about y_{Snow};

this conditional PDF can be written as

$$p(y_{\mathrm{Snow}}|x_{\mathrm{Pa}},x_{\mathrm{Ta}}) = \int_0^\infty p(y_{\mathrm{Snow}}|y_{\mathrm{Pa}},x_{\mathrm{Ta}})\,p(y_{\mathrm{Pa}}|x_{\mathrm{Pa}})\,dy_{\mathrm{Pa}}\,, \tag{3}$$

and this representation suggests that we can split up the statistical model in two separate steps:

1) Generate a forecast distribution $p(y_{\mathrm{Pa}}|x_{\mathrm{Pa}})$ for y_{Pa} using the CSGD-based nonhomogeneous regression method described in section 3b. This forecast distribution can be thought of as removing possible biases from x_{Pa} and representing the associated forecast uncertainty.
2) Use an analog approach to predict y_{Snow} based on the predictors x_{Ta} and y_{Pa} (i.e., based on a predictive sample of *analyzed* precipitation amounts instead of GEFS precipitation forecasts).

Using y_{Pa} instead of x_{Pa} in step 2 has the benefit that the challenging (but crucial) task of predicting precipitation amounts and quantifying the associated uncertainty is handled by a parametric approach with demonstrated good performance in the situation of more extreme events. Finding analog dates separately for the different Pa sample values y_{Pa} corresponding to low, moderate, and heavy precipitation then makes it easier to identify dates with sufficiently similar values for both predictors. Meanwhile, using an analog approach is a convenient way to deal with the highly nonlinear and complex relation between temperature and snowfall amount.

Figure 3 gives a schematic overview over the steps involved in the construction of a 50-member snowfall forecast ensemble based on the predictive CSGD for the analyzed precipitation amounts and the Ta predictor. First, a systematic sample of size 10 is generated as the subset $q_5, q_{10}, \ldots, q_{50}$ of the quantiles $q_1, \ldots, q_{50}$ of today's predictive CSGD with levels $\alpha_k = (k/51),\ k = 1, \ldots, 50$. This sample represents a calibrated probabilistic forecast, but the sample values take the role of the analyzed precipitation amount y_{Pa} in (3). Our particular choice of a 10-member sample is biased toward larger Pa amounts because the 10 quantile levels are not symmetric around 0.5; this is intentional and the rationale behind this choice will be explained below.

The analog method is now applied sequentially: for each value y_{Pa} (starting with the largest) in the predictive sample, five analog dates with similar y_{Pa} and similar x_{Ta} are selected, where similarity is defined as in (1). Multiple selection of the same analog date while cycling through the 10 sampled Pa values is allowed. This potentially reduces the effective analog ensemble size in a data-driven way: if a large training sample is available, different values of y_{Pa} likely result in different analog dates, and the effective ensemble size is close to 50. A small training sample and several unusually large values of y_{Pa}, on the contrary, likely result in the same date being selected multiple times, which reduces the effective ensemble size but avoids a bias toward smaller (more common) values of y_{Pa}. Multiple selection of analog dates can be particularly beneficial at very dry locations where, for a wet forecast, there may not be a sufficient number of historic dates where analyzed Pa amounts were nonzero. Where possible, we only use historic dates with at least 1-mm analyzed Pa as analog candidate dates for a sample value $y_{\mathrm{Pa}} > 0$, since for very light precipitation events the SLRs implied by comparing the analyzed snowfall and precipitation amounts become less and less dependable. At very dry locations, the wettest 10 dates are considered as analog candidate dates, even if some of them have Pa less than 1 mm. By allowing multiple selection of analog dates we can still always assemble a 50-member snowfall ensemble.

Even with the sequential analog search procedure outlined above there is no guarantee that the 50-member Pa ensemble corresponding to the selected analog dates is a good representation of the predictive distribution for Pa. We can further improve the quality of the snowfall ensemble by deriving an adjustment factor f_k for each analog ensemble member k that modifies the precipitation amount a_k such that the ensemble of adjusted values $\tilde{a}_k := f_k a_k$ yields a perfect representation of the predictive CSGD. Assume that the values $a_1, \ldots, a_{50}$ are in ascending order (otherwise reorder the dates). We know that $q_1, \ldots, q_{50}$ is an optimal representation of the calibrated predictive distribution for Pa, and defining $f_k := q_k/a_k$ therefore yields an optimal adjustment of precipitation amounts. What is the appropriate adjustment for the corresponding snowfall amount s_k? If $q_k = 0$, we simply set $\tilde{s}_k := 0$. Otherwise (i.e., if $q_k > 0$), we most likely have $a_k > 0$ due to the wet bias hat was intentionally introduced into the analog search; without the bias, one would have to deal with situations where some $a_k = 0$ needs to be mapped to $q_k > 0$, and a multiplicative adjustment as suggested above is not possible. Our discussion of Fig. 2a suggests a less than proportional increase of snowfall amounts with increasing amounts of SWE, and simply defining $\tilde{s}_k := f_k s_k$ may therefore not be appropriate. Some exploratory analysis shows that for $y_{\mathrm{Pa}} > 1\,\mathrm{mm}$, $y_{\mathrm{Snow}} \sim y_{\mathrm{Pa}}^{0.75}$ is a good approximation for the relationship between snowfall and SWE amounts at most analysis grid points, and we therefore define the adjusted snowfall amount $\tilde{s}_k$ as

$$\tilde{s}_k := \begin{cases} s_k & \text{for} \quad a_k \le 1\,\mathrm{mm} \\ f_k^{0.75} s_k & \text{for} \quad a_k > 1\,\mathrm{mm} \end{cases}. \tag{4}$$

Note that this also covers the case where the precipitation type is rain, and $\tilde{s}_k = s_k = 0$. A beneficial side effect of this adjustment is that snowfall amounts associated with the same analog date are mapped to different values because they are matched to different predictive Pa quantiles. This way, they have the same precipitation type and SLR, but the adjustment still increases the diversity in the ensemble and partially offsets the decrease in effective ensemble size due to multiple selection of analog dates. The adjustment also removes the wet bias introduced through the asymmetric sampling of the predictive CSGD.

The schematic illustration in Fig. 3 is based on a case with rather large forecast precipitation amounts and exemplifies the advantages of the proposed hybrid scheme over the standard analog approach described in section 3a. It is clear that it is easier to find appropriate analog dates for the individual values of y_{Pa} (which sample the entire range of possible predicted outcomes for Pa) since only five dates are required for each sample value. Even then it is difficult to find analog dates for the larger values of y_{Pa}, and this is where the adjustment yields the biggest benefit: it increases some of the larger analog ensemble members, thus generating an adjusted ensemble where Pa values represent the calibrated forecast distribution including the tail with more extreme precipitation amounts. At the same time it separates the Pa (and snowfall) values associated with the same analog date.

Based on the adjusted ensemble of snowfall amounts, probabilistic forecasts such as probabilities of threshold exceedance or quantile forecasts and prediction intervals can be derived. Figure 4 shows 12–36-h ahead quantile forecast (for levels 0.1, 0.5, and 0.9) of snowfall amounts over the mountain region and the corresponding analyzed field. The forecasts are for the example date discussed in Fig. 1 and exemplify the uncertainty about the precipitation type and SLR that comes on top of the forecast uncertainty about precipitation amounts. The underlying GEFS forecasts have largely missed the heavy precipitation over the Uinta Mountains in northeastern Utah, but they provided a sufficiently strong signal for precipitation over southeastern Wyoming and northern Colorado. The uncertainty about temperature, however, results in wide prediction intervals, especially over southeastern Wyoming where conditions are right at the border between rain and snow, and thus various scenarios from almost no snow to rather high snow accumulations are possible.

An example of a northeast coastal snowstorm and associated 12–36-h ahead quantile forecast (this time for levels 0.25 and 0.75) is shown in Fig. 5. Even at these shorter lead times the forecast of the area of most intense snowfall is somewhat displaced: heavy snowfall is predicted for southeastern Massachusetts, while the band of heavy snowfall that actually occurred further inland is not covered by the central 50% prediction interval. At longer lead times (not shown) the forecast signal was rather weak everywhere in this domain, and the observed event falls in the tail of the distribution. This is another reminder of the difficulties associated with snowfall prediction, and it raises the question of how much skill one can expect from these probabilistic forecasts. This will be studied systematically in the next section.

4. Verification of the probabilistic snowfall predictions

We compare probabilistic, GEFS-based snowfall forecasts generated by the standard analog method described in section 3a and by the hybrid parametric/analog scheme proposed in section 3c. Each month is processed separately, and only the months November–April where snowfall is most common in the two study areas are considered for verification. The four cool seasons for

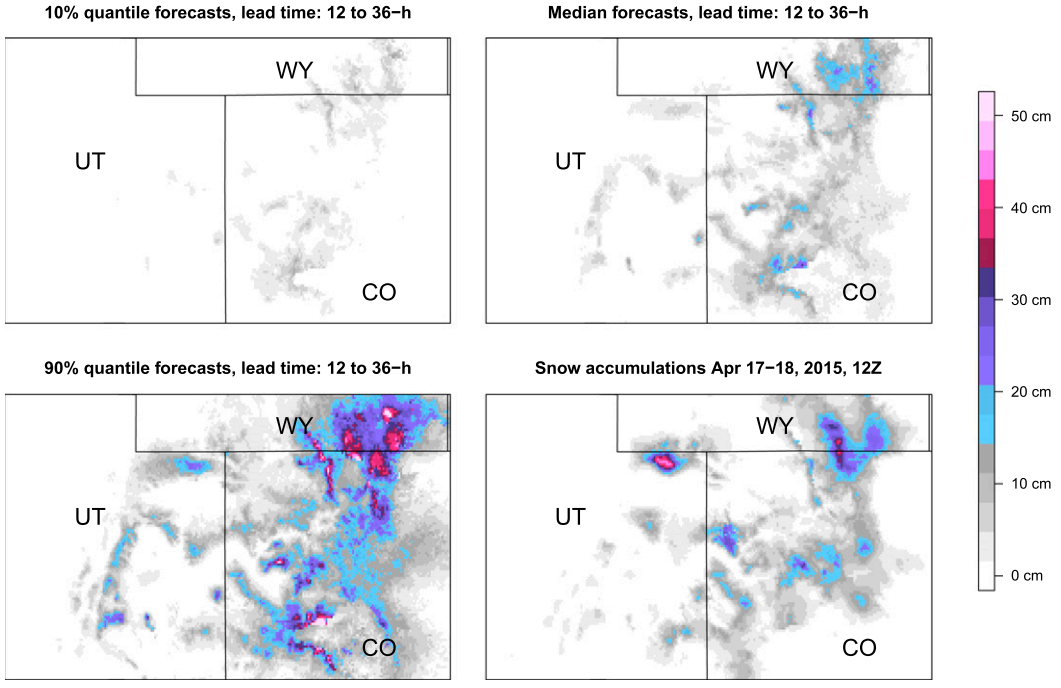

FIG. 4. Analyzed snowfall amounts over the mountain region for the 24-h period ending at 1200 UTC 18 Apr 2015, and GEFS-based, 1-day-ahead quantile forecasts obtained with the hybrid parametric–analog scheme described above.

which forecast and observation data are available (2013/14, . . . , 2016/17) are cross validated (i.e., one cool season is held out for validation), probabilistic forecasts are generated based on training data from the three remaining cool seasons, and the held-out season is cycled through so that we obtain verification results for all four cool seasons that were obtained with independent training and validation datasets.

a. Sensitivity to tuning parameters

The performance of the analog method for snowfall amounts described in section 3a depends on two tuning parameters: the number of selected analog ensemble members n_{ana} and the weight w_{Pa} in the dissimilarity measure Δ_d in (1). The other weight was fixed at $w_{\text{Ta}} = 1$ without loss of generality since only the relative weight of the different predictors matters. In this subsection we study the impact of different choices of n_{ana} and w_{Pa} on the quality of the resulting snowfall analog ensemble. The same two tuning parameters are also part of the analog component of the proposed hybrid parametric/analog scheme. In our description of this method in section 3c we have assumed $n_{\text{ana}} = 50$ and pointed out that the effective analog ensemble size can be smaller because the proposed algorithm permits multiple selection of analog dates. However, we can still vary n_{ana} in this

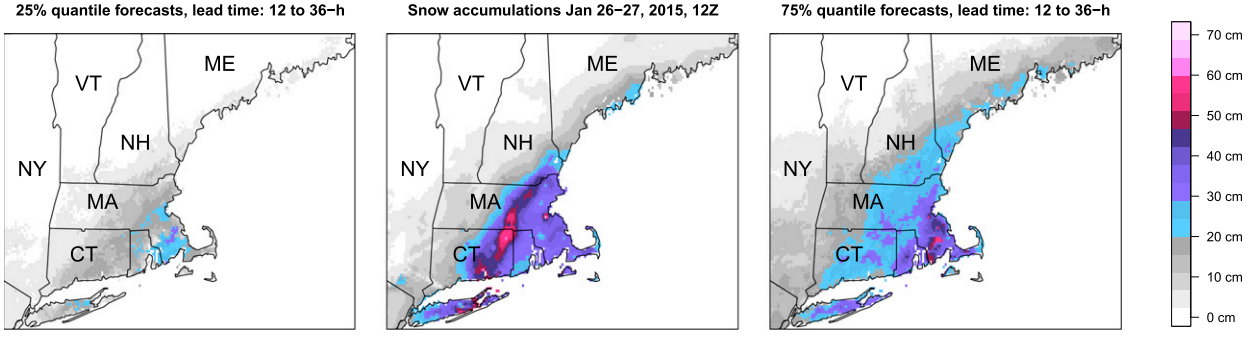

FIG. 5. Analyzed snowfall amounts over the northeast region for the 24-h period ending at 1200 UTC 27 Jan 2015, and GEFS-based, 1-day-ahead quantile forecasts obtained with the hybrid parametric–analog scheme described above.

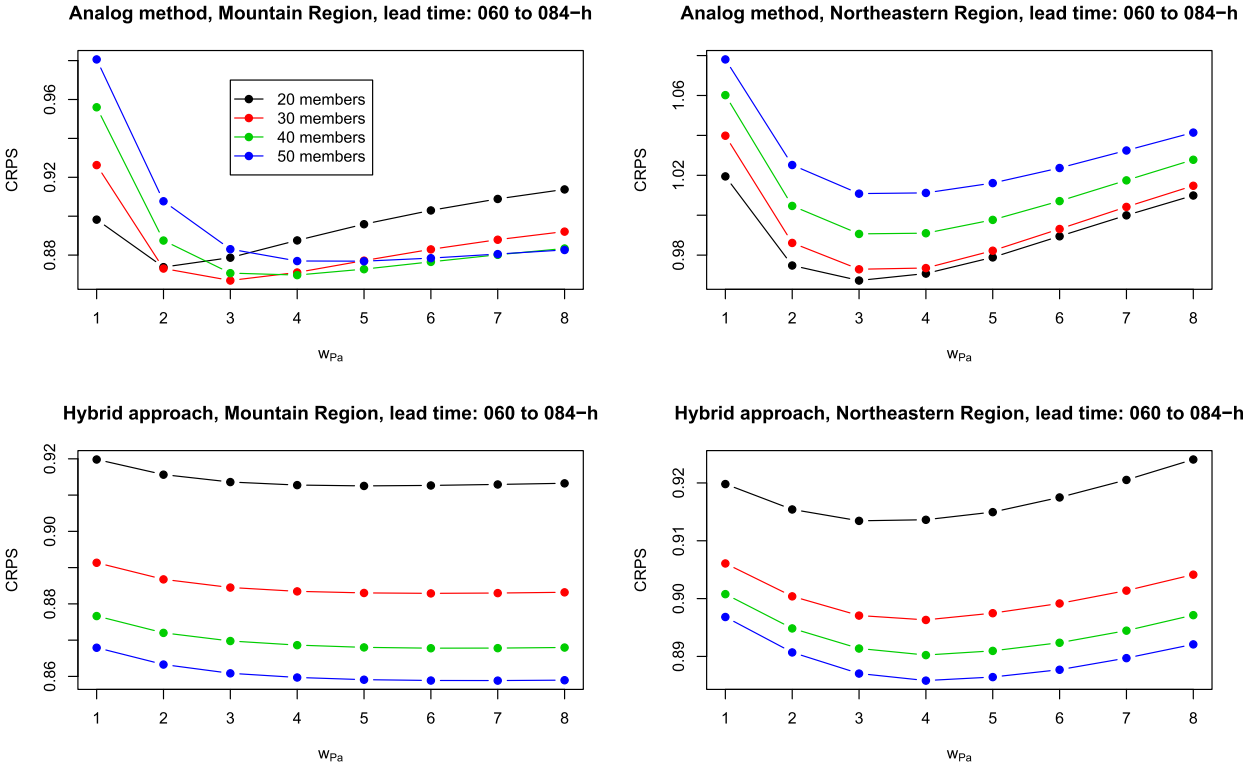

FIG. 6. CRPSs for the standard analog method and the hybrid parametric–analog approach using different ensemble sizes and weights for the Pa predictor. The scores shown here are averages over all grid points in the respective region and all days in January during the four cool seasons considered in this study.

algorithm and reduce the number of quantiles used to represent the predictive distribution of precipitation amounts accordingly. We assess the predictive performance of the snowfall ensembles generated with the two different methods and different values of n_{ana} and w_{Pa} by using the sample version [Grimit et al. 2006, their Eq. (3)] of the continuous ranked probability score (CRPS), that is,

$$\text{CRPS}(\mathbf{x}, y) = \frac{1}{n_{\text{ana}}} \sum_{i=1}^{n_{\text{ana}}} |x_i - y| - \frac{1}{2n_{\text{ana}}^2} \sum_{i=1}^{n_{\text{ana}}} \sum_{j=1}^{n_{\text{ana}}} |x_i - x_j|, \quad (5)$$

where $\mathbf{x} = (x_1, \ldots, x_{n_{\text{ana}}})$ are the ensemble member forecasts and y is the verifying observation, and a lower score means better performance. The CRPS is a common measure for the overall performance of probabilistic forecasts that takes both reliability ("Does the ensemble provide an accurate representation of forecast uncertainty?") and sharpness ("Is the ensemble spread as small as possible, given reliability?") into account.

Figure 6 shows the average CRPS as a function of w_{Pa} and n_{ana} where averages are taken over all grid points in the respective region and all days in January during the four cool seasons considered in this study. Only results for the forecast lead time period 60–84 h ahead are shown here (additional plots for different forecast lead times and different months are provided in online supplemental material A). The conclusions about the optimal analog ensemble size for our implementation of an analog method for snowfall amounts are in line with those reported by Hamill et al. (2006) for precipitation amounts in that the optimal n_{ana} increases with forecast lead time. This is because larger ensembles reduce the sampling variability, but this comes at the expense of selecting less similar analog dates. At short forecast lead times where the GEFS forecasts of the predictors Pa and Ta have relatively good skill, a high degree of similarity of today's forecasts and the forecasts at the analog dates is very important; at longer forecast lead times where the GEFS forecast skill is lower, the optimal trade-off shifts toward a reduction in sampling variability even if that goes along with a lower degree of similarity of some analog dates. This trade-off is very different for the hybrid method proposed in section 3c where a reduction in the ensemble size always results in lower performance compared to the proposed choice of $n_{\text{ana}} = 50$. As explained above, this is because the effective ensemble size for this method is

often smaller than n_{ana}, especially if more extreme precipitation events are forecast by the GEFS.

With regard to the weight coefficient we find that the snowfall analog method performs best with values $w_{Pa} \approx 3$ in January and $w_{Pa} \approx 2$ in April. A possible explanation of this dependence on the season is that in winter it is very likely that the precipitation type is snow at most locations, and so the forecast precipitation amount is more important for predicting snow accumulations. In spring and fall, on the contrary, the determination of precipitation type plays a bigger role for predicting the range of outcomes for snow accumulations, and therefore the temperature predictor carries somewhat more weight. Sensitivity to the choice of w_{Pa} is rather low for the hybrid parametric/analog scheme, but generally higher values $w_{Pa} \approx 4$ or $w_{Pa} \approx 5$ are preferred. This is likely because Δ_d is now based on predictive samples of analyzed precipitation. These samples already represent a range of possible outcomes, and so each individual sample value should be matched more closely.

Based on the conclusions drawn from Fig. 6 (and the additional plots provided in supplemental material A) we make the following parameter choices for our analog method for the subsequent analyses: we use $w_{Pa} = 3$ during the winter months and $w_{Pa} = 2$ during spring and fall. Further, we let the ensemble size increase with forecast lead time and use $n_{ana} = 15 + 5t_{lead}$, where t_{lead} is the forecast lead time in days (e.g., 3 for the lead time period of 60–84 h ahead). For the hybrid parametric/analog scheme we use $n_{ana} = 50$ and $w_{Pa} = 4$ for every month and all forecast lead times. For optimal performance in an operational setting one could select optimized parameters n_{ana} and w_{Pa} for each month and each lead time, possibly even separately for each location if the training sets are sufficiently large so that overfitting is not a concern. A brute force optimization as described by Junk et al. (2015) could be performed to make that selection without the need for any user interaction. Our parameter choices motivated by the above sensitivity analysis are not necessarily the optimal choice in every single situation, and the scores obtained with them might differ from the scores that could be obtained via brute force optimization of the tuning parameters. However, we expect the results obtained with our parameter choices to be close to optimal and robust against overfitting.

b. Reliability

One requirement for probabilistic forecasts to be useful is reliability, the property that an event predicted to occur with probability p occurs with a relative frequency $\approx p$ when all verifying observations associated with forecast probabilities $\approx p$ are considered. Here, we study the events of snow accumulations exceeding 1 cm, 10 cm, and 25 cm, respectively, for forecast lead times of 1 day, 3 days, and 5 days ahead. Reliability diagrams for forecasts obtained with the hybrid parametric/analog method with cases pooled across all grid points within our study area, all four verification years and all six cool season months are shown in Figs. 7 and 8 (reliability diagrams for the analog method are provided in supplemental material B). Over the mountain region, the probability forecasts derived from the postprocessed snowfall ensembles are almost perfectly reliable for lead times of 1 and 3 days, but for 5 days of forecast lead time they become slightly overconfident. A look at the reliability diagrams for precipitation amounts (not shown here) reveals that the overconfidence of the snowfall forecasts is inherited from the postprocessed precipitation forecasts; the overconfidence of those can in turn be explained by the degradation of the signal-to-noise ratio of the raw GEFS forecasts in combination with relatively small training samples compared to other applications in which the CSGD-based nonhomogeneous regression method was used. Over the northeast region, reliability is not quite as good with reliability curves for several threshold and forecast lead time suggesting that the underlying exceedance probability forecasts are somewhat underconfident. Except for the 1-cm threshold at 12–36-h forecast lead time, however, departures from the diagonal are mostly within the boundaries compatible with sampling variability.

c. Brier skill scores

With reliability being established, the next important validation aspect concerns the skill of the probabilistic, GEFS-based snowfall forecasts relative to climatology. Considering again probability forecasts of threshold exceedance, we study Brier skills scores of both standard analog ensemble forecasts and hybrid parametric/analog ensemble forecasts. Denote by $p_{an,i}$ and $p_{hy,i}$ the respective probability forecasts for verification case i, and by $p_{cl,i}$ the corresponding climatological relative frequency of exceedance. The latter were calculated separately for each analysis grid point and each month, using snowfall analysis data from ± 45 days around the 15th of the respective month during all four cool seasons. Apart from the fact that we did not cross validate the data used to calculate these climatological exceedance probabilities, this way of composing these datasets is similar to how the training data for the different postprocessing methods were composed; the underlying idea is again that sufficiently large datasets are required to obtain stable estimates of climatological exceedance probabilities, while a reasonable climatology needs to account for spatial and seasonal variability.

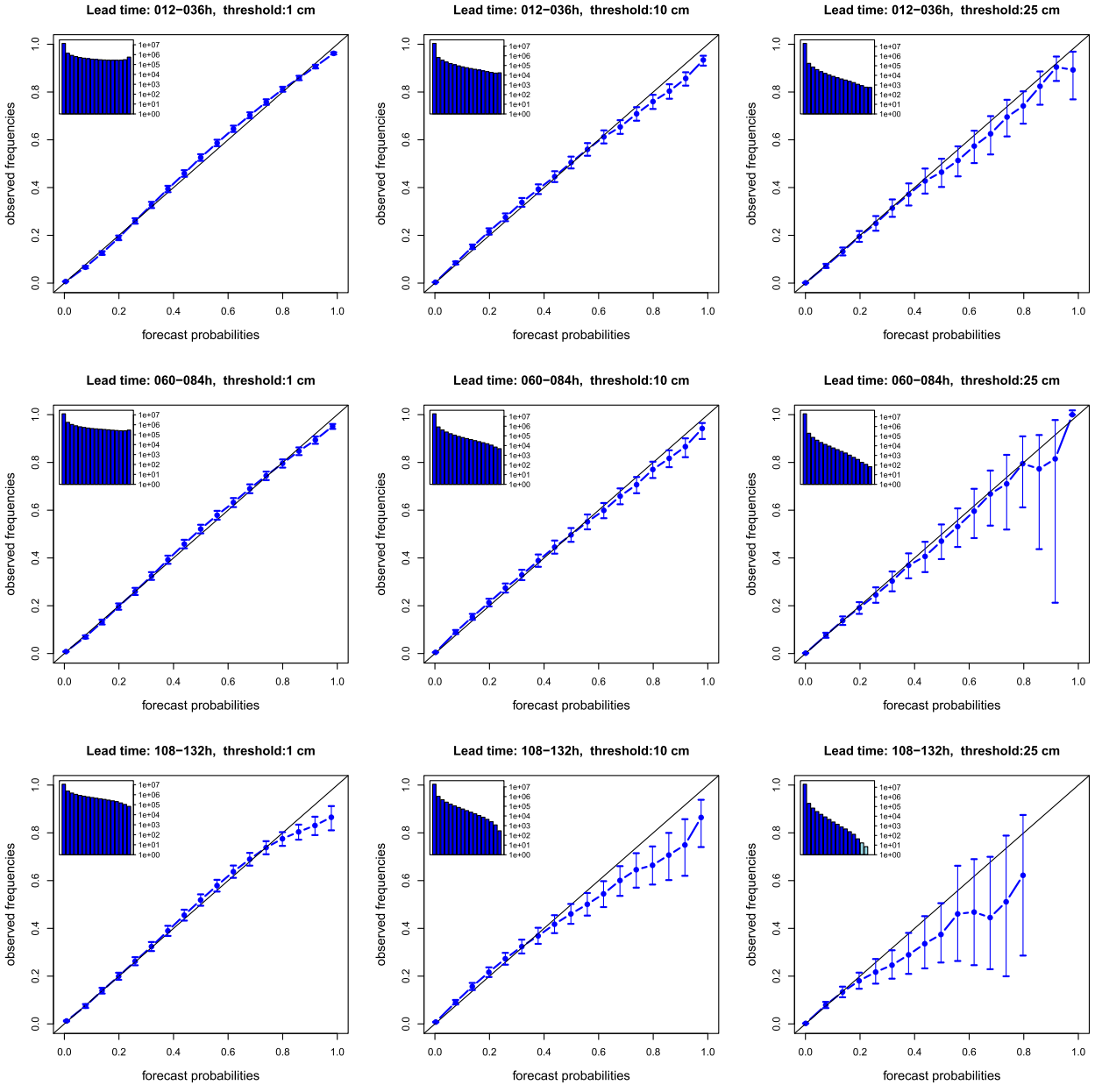

FIG. 7. Reliability diagrams for probability forecasts obtained with the hybrid parametric–analog scheme over the mountain region. The inset histograms depict the frequency with which the different forecast probabilities were issued. The vertical bars represent 90% confidence intervals obtained by bootstrap resampling.

Denote by o_i the verifying binary observation (i.e., one for exceedance, zero for nonexceedance). We then calculate the mean Brier scores over N cases (all analysis grid points within the respective study area times all four cool seasons):

$$\overline{\mathrm{BS}}_{\star} = \frac{1}{N}\sum_{i=1}^{N}(o_i - p_{\star,i})^2, \tag{6}$$

where $\star$ is either "an," "hy," or "cl," and depict the Brier skill scores

$$\overline{\mathrm{BSS}}_{\star} = 1 - \frac{\overline{\mathrm{BS}}_{\star}}{\overline{\mathrm{BS}}_{\mathrm{cl}}} \tag{7}$$

for the analog and the hybrid method in Fig. 9. Relating the Brier scores of the two postprocessing schemes to climatological scores allows one to judge how much information beyond simple climatological averages the respective forecasts can provide, with the maximum skill being 1 and values less or equal to 0 suggesting that the method provides probabilistic guidance that does not

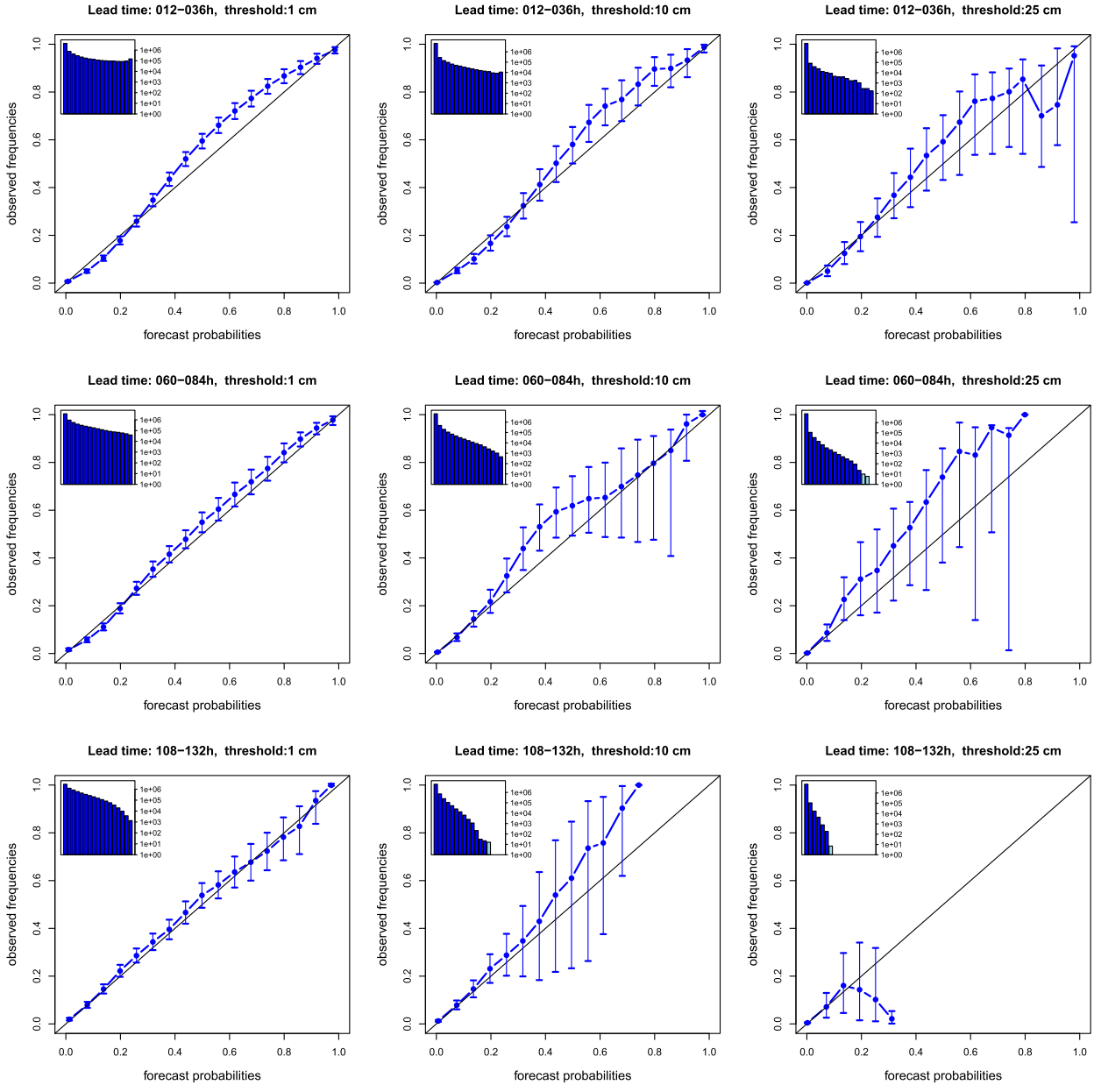

FIG. 8. As in Fig. 7, but for the northeast region.

add any value over climatological guidance. To assess whether the Brier score differences between the two methods are statistically significant, we perform one-sided Wilcoxon signed-rank tests (Wilks 2011, chapter 5.3) of the null hypothesis that the Brier scores obtained with the analog method are lower (i.e., better) or equal to those obtained with the hybrid parametric/analog scheme. Following Hamill (1999) we assume independence between the scores associated with different days, but we account for spatial dependence by applying the statistical tests to average Brier scores over the respective domains. If we can reject the null hypothesis at the 5% level, we will say that the forecasts obtained with the hybrid scheme are significantly better, and we mark the corresponding month with the symbol "×" in the color of the respective forecast lead time. Figure 9 shows that the proposed hybrid scheme yields probabilistic forecasts that are significantly more skillful than those obtained with the standard analog approach at threshold 1 cm for all forecast lead times, and at the higher thresholds for the forecast lead time 12 to 36 h. For the 25-cm threshold in the

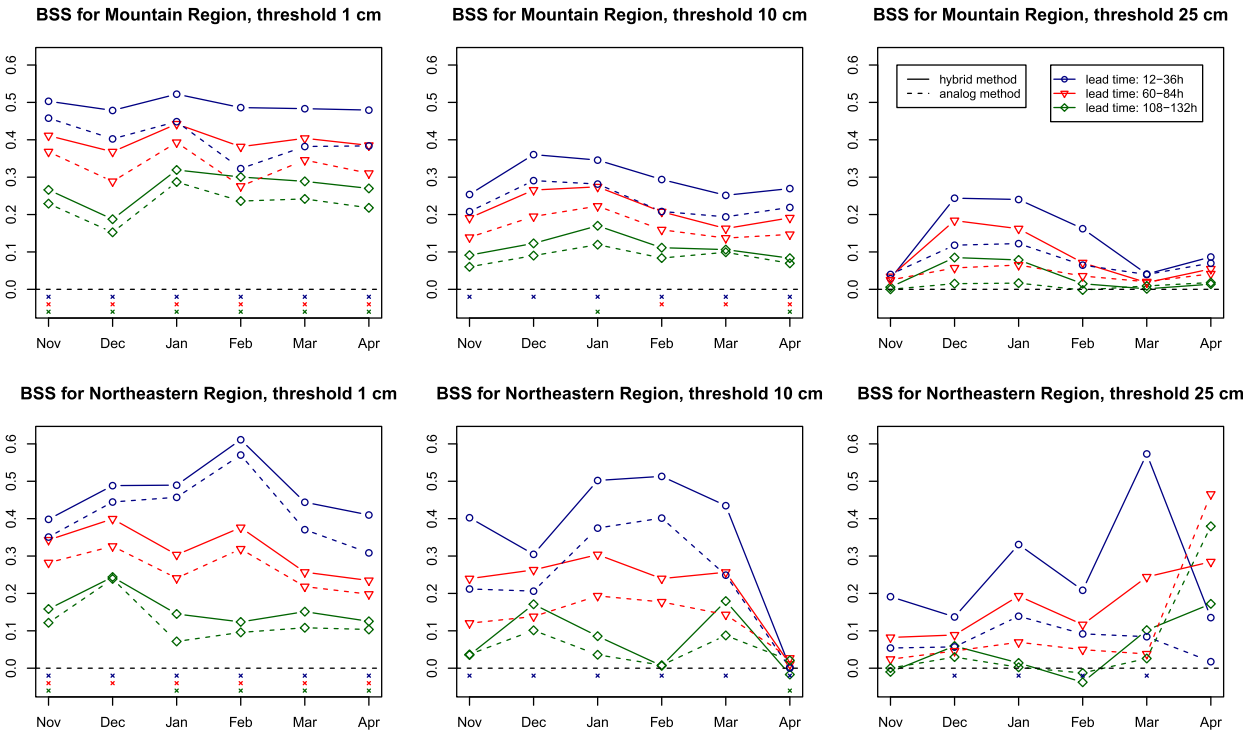

FIG. 9. Brier skill scores for the standard analog method and the hybrid parametric–analog scheme. Statistically significant differences between the two methods are marked with an "×."

mountain region, the Brier skill scores obtained with the hybrid scheme are better for all months and all lead times, but the differences are not statistically significant. For both methods, skill drops rapidly with forecast lead time as a result of the combined forecast uncertainty about precipitation amounts and temperature, which makes prediction at longer leads rather challenging, especially when the focus is on heavy snowfall events.

d. *Mean absolute error and interval skill scores*

Quantile forecasts are another quantity of interest that can be derived from either standard analog or hybrid parametric/analog ensemble. The ensemble median can be considered as a deterministic snowfall forecast. It is the optimal point forecast with respect to the mean absolute error (Gneiting 2011), and we therefore measure its quality by the mean absolute error (MAE) skill score with respect to the climatological median, calculated again from analyzed data from ±45 days around the 15th of the respective month. In addition, we study $(1-\alpha)\times 100\%$ central prediction intervals for $\alpha = 0.2$ and $\alpha = 0.5$, which are defined by the forecast quantiles (q_{10}, q_{90}) and (q_{25}, q_{75}), respectively. A suitable performance measure for these prediction intervals is the skill score associated with the *interval score*

$$\overline{\mathrm{IS}}_{\star} = \frac{1}{N}\sum_{i=1}^{N}(u_i - l_i) + \frac{2}{\alpha}(l_i - y_i)\mathbf{1}_{\{y_i<l_i\}} + \frac{2}{\alpha}(y_i - u_i)\mathbf{1}_{\{y_i>u_i\}} \quad (8)$$

[see Gneiting and Raftery (2007) and references therein] where y denotes the verifying analysis, and l, u are the lower and upper bound of the prediction interval, respectively. As in section 4c we use one-sided Wilcoxon signed-rank tests to test whether the scores obtained with the hybrid approach are significantly better. The results depicted in Fig. 10 confirm our previous finding that snowfall predictions obtained with the hybrid parametric/analog scheme are more skillful than those obtained with the standard analog scheme. All skill differences between the two methods are statistically significant for forecast lead time 12–36 h, most of them are significant for forecast lead time 60–84 h, and for forecast lead time 108–132 h most of the score differences in the mountain region and about half of the score differences in the northeastern region are statistically significant. Deterministic forecast skill drops rapidly with forecast lead time, especially in the northeast region, but the probabilistic information encoded in the prediction intervals can still add value over climatology, even though the prediction intervals can

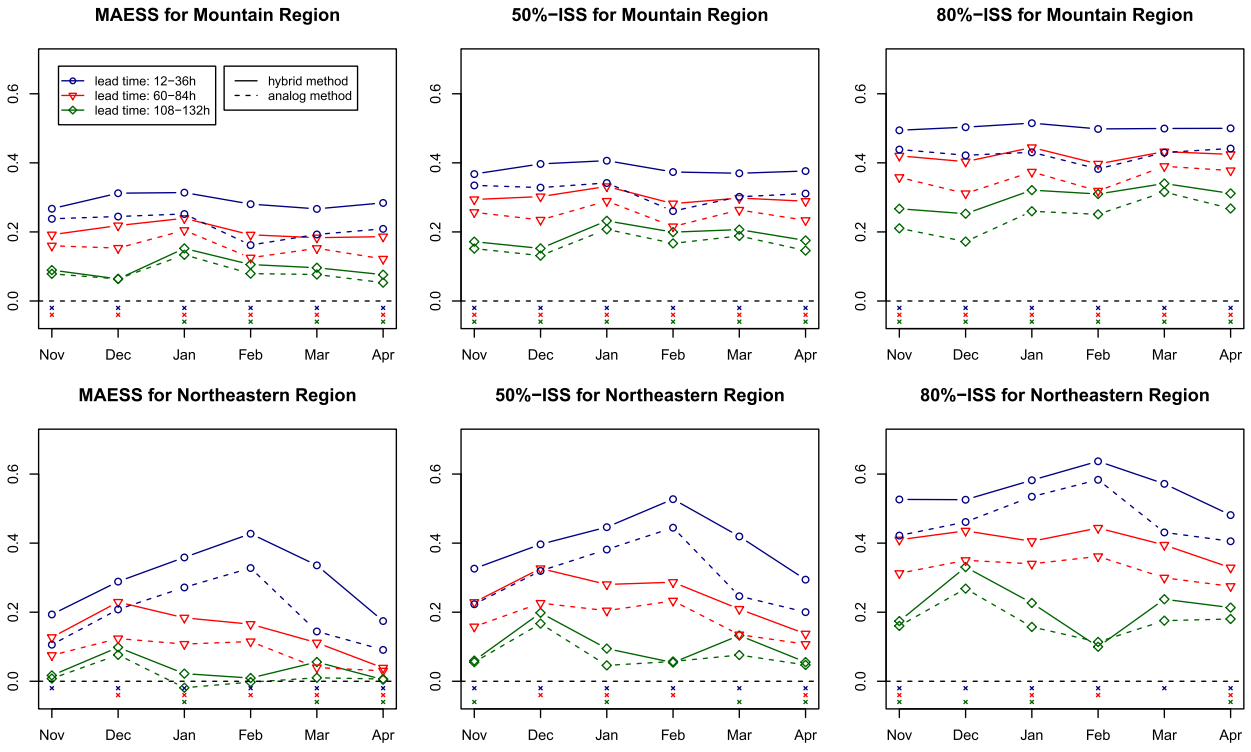

FIG. 10. Mean absolute error and interval skill scores for 50% and 80% prediction intervals for the standard analog method and the hybrid parametric–analog scheme. Statistically significant differences between the two methods are marked with an "×."

sometimes be very wide as discussed in the context of the example in Fig. 4.

e. *Use of additional predictors*

So far, Ta (maximum wet-bulb temperature aloft) was the only predictor considered in order to determine how precipitation amounts translate into snowfall amounts. As mentioned in the introduction and demonstrated in the literature (Ware et al. 2006; Alcott and Steenburgh 2010), there are a number of other variables such as wind speed or relative humidity that affect the snow-forming process and therefore the SLR. While precipitation amount and temperature arguably play the most important role for determining the snowfall amount, it is worth studying if the results reported above can be improved by including additional predictors related to SLR in the analog component of the hybrid parametric/analog scheme proposed in section 3c. Additional predictors may improve the (implicit) estimation of SLRs associated with the snowfall ensemble forecasts, but they could also have a negative impact since with every additional predictor variable it becomes more challenging to find an analog date with similar values across all of these predictors; a weather variable that is only weakly related to snowfall amounts or is not well predicted by medium-range weather prediction systems may introduce additional sampling variability in the analog search and thus do more harm than good.

We consider two additional predictors already described in section 2: 80-m wind speed forecasts and 2-m relative humidity forecasts. To extend the hybrid scheme proposed in section 3c we just need to add an additional term $w_{\mathrm{w80}}|\mathrm{W80}^d - \mathrm{W80}^t|$ or $w_{\mathrm{RH}}|\mathrm{RH}^d - \mathrm{RH}^t|$ to the dissimilarity measure Δ_d in (1). For the month of January, Fig. 11 shows continuous ranked probability skill scores (CRPSSs) for different values of w_{w80} and w_{RH}; the weights w_{Ta} and w_{Pa} are kept fixed at the values determined in section 3a. Here, skill is calculated relative to the hybrid ensemble forecast based on Pa and Ta only. One-sided Wilcoxon tests are used to test whether the decrease in CRPS obtained with the best possible weight coefficient is statistically significant, the resulting p values are printed into Fig. 11 next to the associated CRPSS values. Based on these results, we conclude that neither 80-m wind speed forecasts nor 2-m relative humidity forecasts yield a significant skill improvement when included as additional predictors in our hybrid scheme. The uncertainty associated with their prediction is too large relative to their information content about SLR to make use of that information. This conclusion can of course be different for different regions and/or different forecast systems which could have more

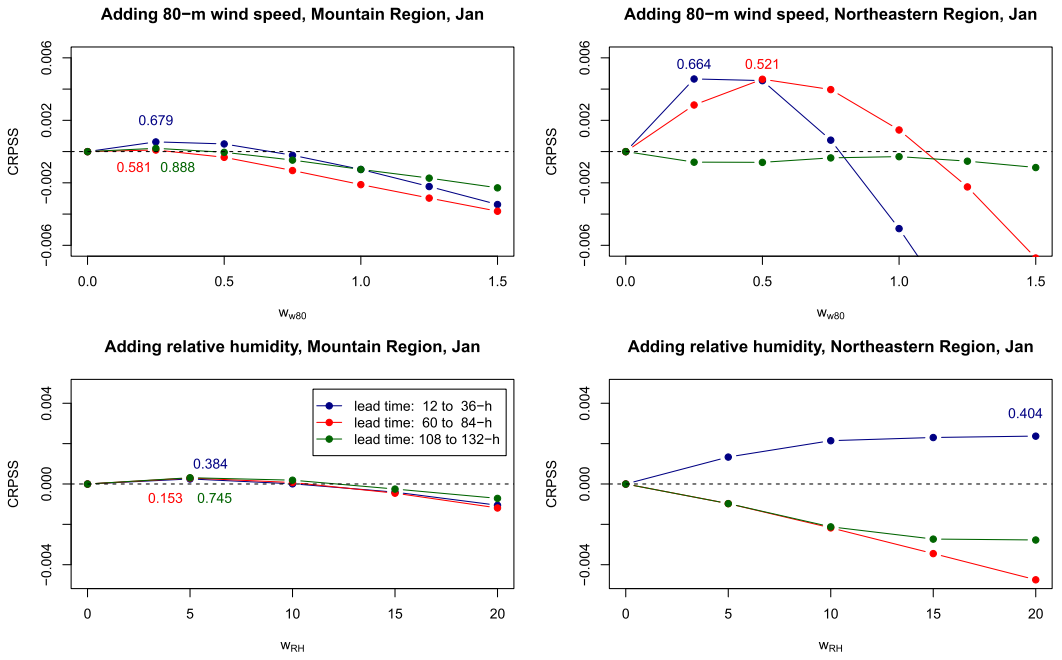

FIG. 11. CRPSSs of the snowfall ensemble forecasts obtained with the hybrid parametric–analog scheme with added W80 or RH predictor, relative to the scores obtained with just the Ta predictor. The numbers are the *p* values obtained from a one-sided Wilcoxon signed-rank test for significant reduction of CRPS.

favorable error characteristics with regard to these variables. It might also help to use related but different predictors that are not subject to the operational constraints that we imposed on ourselves in order to be compatible with NBM weather elements.

5. Discussion

We have proposed a hybrid scheme for probabilistic snowfall forecasting that employs the CSGD-based regression method to generate reliable predictive distributions for precipitation amounts, and then uses samples from these predictive distributions along with GEFS ensemble mean forecasts of maximum wet-bulb temperature aloft as predictors in an analog scheme. Our particular choice of predictors was based on existing literature (Ware et al. 2006; Alcott and Steenburgh 2010; T. Alcott 2017, personal communication), and availability of these predictors as weather elements in NOAA's National Blend of Models (NBM) to which the proposed algorithm might be added in the future. The proposed hybrid method was demonstrated over an area in the mountain west of the United States and a coastal area in the northeastern United States, and verification results over four cross-validated cool seasons suggest that the resulting snowfall forecasts are more skillful than those obtained with the standard implementation of an analog method.

The proposed hybrid scheme leverages the flexibility of an analog method in accounting for complex forms of nonlinearity and heteroscedasticity in the predictor–predictand relationships and the ability of a parametric postprocessing method for precipitation amounts to extrapolate predictor–predictand relationships from more common events to rare events. In areas where snowfall is possible but uncommon, even that strategy might fail to identify a sufficient number of analog cases, and might have to be complemented with concepts like *supplemental locations* (Hamill et al. 2015; Lerch and Baran 2017), where a set of additional locations with similar climatological and terrain characteristics are identified for each location, and the respective data at these supplemental locations is used to augment the training dataset at the location of interest.

The focus of this paper was on predicting snowfall amounts. Related quantities of interest like probability of snow (vs rain) or snow-to-liquid ratio are predicted implicitly, and could partly be reconstructed from the snowfall ensembles generated by either analog or hybrid method, but neither of the two techniques has been optimized for these quantities. Probabilistic forecasting approaches have been proposed that target precipitation type (Scheuerer et al. 2017a) and snow-to-liquid ratio (Roebber et al. 2003) directly, and are more appropriate for that purpose.

Both methods studied in this paper were demonstrated with 24-h snowfall accumulations, since this is the temporal resolution for which the NOHRSC snowfall analyses are currently available. Our strategy of weighting the 6-h temperature and wind speed forecasts proportionally to the corresponding precipitation amounts emphasizes the values of these predictors during the period where most of the precipitation occurs. However, it cannot explicitly account for temperature variations—and thus changes in precipitation type or SLR—during the 24-h forecast periods. In the future, NOHRSC snowfall analyses at a 6-h temporal resolution will likely become available, and this would mitigate this issue and also allow us to address the NBM demand for 6-h snow forecasts without having to use the ensemble copula coupling technique for temporal disaggregation as proposed by Stauffer et al. (2018).

An algorithm for statistical postprocessing of precipitation amounts that is tailored to the requirements of NBM has been developed (Hamill and Scheuerer 2018) and will soon be transferred to operations. The parametric/analog scheme for probabilistic forecasting of snowfall amounts proposed here can be adapted and used in combination with this technique, and thus become part of NBM itself. For many of the forecast systems used in the NBM reforecast data are not available, which poses some challenges since biases (e.g., in temperature forecasts) may change with model version changes. Since deterministic forecasts of maximum wet-bulb temperature aloft are a weather element in the NBM, however, one could simply use these (bias corrected) NBM forecasts in (1) instead of the raw ensemble mean and thereby avoid the implications of model version changes. For the snowfall and precipitation analyses used in our algorithm, on the contrary, it is vital to have a record of as many years of data as possible to ensure that the analog search can draw from a wide range of different weather situations.

Acknowledgments. The authors thank Trevor Alcott for helpful discussions concerning the meteorological background of the snow-forming process and the choice of suitable predictors that can be used in a postprocessing algorithm. We thank Greg Fall for explaining the details about the generation of the NOHRSC snowfall analyses, and we are grateful for many constructive comments by the three anonymous reviewers that helped improve the presentation of this article. Our research was supported by a grant from the NOAA/NWS Research to Operations (R2O) initiative for the Next-Generation Global Prediction System (NGGPS), Award NA15OAR4320137, and by a funding agreement between NOAA/ESRL/PSD and NOAA/NWS/MDL on the development and transfer of advanced statistically postprocessed probabilistic ensemble guidance.

REFERENCES

Alcott, T. I., and W. J. Steenburgh, 2010: Snow-to-liquid ratio variability and prediction at a high-elevation site in Utah's Wasatch Mountains. *Wea. Forecasting*, **25**, 323–337, https://doi.org/10.1175/2009WAF2222311.1.

Clark, E. P., 2017: Updated NWS Technical Implementation Notice 15-05. NOAA/NWS, accessed December 2017, http://www.nws.noaa.gov/os/notification/tin15-05bigrsc_snowfall_aaa.htm.

Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quart. J. Roy. Meteor. Soc.*, **143**, 909–916, https://doi.org/10.1002/qj.2975.

Delle Monache, L., T. Eckel, D. Rife, and B. Nagarajan, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Gneiting, T., 2011: Making and evaluating point forecasts. *J. Amer. Stat. Assoc.*, **106**, 746–762, https://doi.org/10.1198/jasa.2011.r10138.

——, and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

Gowan, T. M., W. J. Steenburgh, and C. S. Schwartz, 2018: Validation of mountain precipitation forecasts from the convection-permitting NCAR ensemble and operational forecast systems over the western United States. *Wea. Forecasting*, **33**, 739–765, https://doi.org/10.1175/WAF-D-17-0144.1.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.*, **132**, 2925–2942, https://doi.org/10.1256/qj.05.235.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, https://doi.org/10.1175/BAMS-87-1-33.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea.*

Rev., **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Junk, C., L. Delle Monache, S. Alessandrini, G. Cervone, and L. Bremen, 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteor. Z.*, **24**, 361–379, https://doi.org/10.1127/metz/2015/0659.

Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc. C*, **66**, 29–51, https://doi.org/10.1111/rssc.12153.

Lin, Y., 2011: GCIP/EOP surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV data (version 1.0). NCAR/UCAR, accessed December 2017, https://doi.org/10.5065/d6pg1qdd.

R Core Team, 2017: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Roebber, P. J., S. L. Bruening, D. M. Schultz, and J. V. Cortinas Jr., 2003: Improving snowfall forecasting by diagnosing snow density. *Wea. Forecasting*, **18**, 264–287, https://doi.org/10.1175/1520-0434(2003)018<0264:ISFBDS>2.0.CO;2.

Rogers, E., and Coauthors, 2017: Mesoscale modeling development at the National Centers for Environmental Prediction: Version 4 of the NAM forecast system and scenarios for the evolution to a high-resolution ensemble forecast system. *28th Conf. on Weather and Forecasting/24th Conf. on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., 3B.4, https://ams.confex.com/ams/97Annual/webprogram/Paper311212.html.

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, https://doi.org/10.1175/MWR-D-15-0061.1.

——, and ——, 2018: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *J. Hydrometeor.*, **19**, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1.

——, S. Gregory, T. M. Hamill, and P. E. Shafer, 2017a: Probabilistic precipitation type forecasting based on GEFS ensemble forecasts of vertical temperature profiles. *Mon. Wea. Rev.*, **145**, 1401–1412, https://doi.org/10.1175/MWR-D-16-0321.1.

——, T. M. Hamill, B. Whitin, M. He, and A. Henkel, 2017b: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatio-temporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

Sievers, O., K. Fraedrich, and C. Raible, 2000: Improving snowfall forecasting by accounting for the climatological variability of snow density. *Wea. Forecasting*, **15**, 623–629, https://doi.org/10.1175/1520-0434(2000)015<0623:SAAEPO>2.0.CO;2.

Stauffer, R., G. J. Mayr, J. W. Messner, and A. Zeileis, 2018: Hourly probabilistic snow forecasts over complex terrain: A hybrid ensemble postprocessing approach. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **4**, 65–86, https://doi.org/10.5194/ascmo-4-65-2018.

Ware, E. C., D. M. Schultz, H. E. Brooks, P. J. Roebber, and S. L. Bruening, 2006: Improving snowfall forecasting by accounting for the climatological variability of snow density. *Wea. Forecasting*, **21**, 94–103, https://doi.org/10.1175/WAF903.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Academic Press, 704 pp.

Zhao, Q., and F. H. Carr, 1997: A prognostic cloud scheme for operational NWP models. *Mon. Wea. Rev.*, **125**, 1931–1953, https://doi.org/10.1175/1520-0493(1997)125<1931:APCSFO>2.0.CO;2.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.