Comparing Area Probability Forecasts of (Extreme) Local Precipitation Using Parametric and Machine Learning Statistical Postprocessing Methods

KIRIEN WHAN AND MAURICE SCHMEITS

R&D Weather and Climate Modeling, Royal Netherlands Meteorological Institute (KNMI), De Bilt, Netherlands

(Manuscript received 1 November 2017, in final form 31 May 2018)

ABSTRACT

Probabilistic forecasts, which communicate forecast uncertainties, enable users to make better weather-based decisions. Using precipitation and numerous instability indices from the deterministic model HARMONIE–AROME (HA; a nonhydrostatic numerical weather prediction model) as potential predictors, we generate summer areal probabilistic maximum hourly precipitation forecasts across 11 regions of the Netherlands. We compare the skill of three statistical postprocessing methods: an extended logistic regression (ELR), a zero-adjusted gamma distribution (ZAGA), and a machine learning-based method, quantile regression forests (QRF). Forecast skill for low and moderate precipitation thresholds increases with the inclusion of extra predictors, in addition to HA precipitation. HA precipitation is the most important predictor at all lead times in ELR and QRF, while in ZAGA, the most important predictor for the location parameter shifts over lead times from HA precipitation to indices of atmospheric instability. All three methods improve upon a climatological forecast for low and moderate precipitation thresholds. ZAGA and QRF are generally the most skillful methods at moderate thresholds. QRF tends to be the most skillful method at higher thresholds, particularly during the afternoon period. Forecasts are reliable at low and moderate thresholds but tend to be overconfident at higher thresholds. QRF and ZAGA have more potential economic value than the deterministic forecast, with value remaining at high thresholds. A maximum local hourly precipitation threshold of 30 mm h^{-1} (a criterion in the Royal Netherlands Meteorological Institute's code yellow warning for severe thunderstorms) is skillfully forecast by QRF in the afternoon period at short lead times.

1. Introduction

Forecast uncertainty is inevitable, given the chaotic nature of the atmosphere and unavoidable errors, such as those that remain in the initial conditions and the physical parameterization schemes of the numerical model, among others. Deterministic forecasts ignore this uncertainty and leave the user to add their own estimates of confidence to the forecast, while probabilistic forecasts explicitly communicate the uncertainty and enable users to make better decisions (Morss et al. 2008; Joslyn and Savelli 2010; Joslyn and LeClerc 2012; LeClerc and Joslyn 2015).

Global ensemble prediction systems (EPS) now play an essential role in estimating the uncertainties in numerical weather prediction (NWP). Mesoscale EPS are becoming increasingly common, but computing resources still limit the widespread use of high-resolution EPS, particularly in smaller national meteorological agencies. A high-resolution, nonhydrostatic model that explicitly resolves deep convection is essential for forecasting extreme warm-season precipitation events. NWP models that run without a parameterization for deep convection can better represent convective processes; however, errors in the location of precipitation events remain (e.g., Clark et al. 2009; Gagne et al. 2014; Pinto et al. 2015; Herman and Schumacher 2016).

Extreme weather events can have large impacts on society, infrastructure, and the economy. Skillful and reliable forecasts of these events can be beneficial to many facets of the community. Many extreme events, particularly warm-season thunderstorms, occur on small spatial scales. Nonhydrostatic NWP models with high horizontal resolution are necessary to fully resolve these events. Deterministic forecasts from high-resolution models often contain errors in the location of small-scale extreme events and offer no information on the uncertainty in the forecast. Postprocessing of extreme

Corresponding author: Kirien Whan, whan@knmi.nl

DOI: 10.1175/MWR-D-17-0290.1

events is challenging, given their infrequent occurrence, and this makes a large training dataset necessary (e.g., Hamill et al. 2013; Scheuerer and Hamill 2015). The benefits to society of improved forecasts of extreme precipitation events are numerous, making improving these forecasts a challenging but important question. Here, we focus on extreme precipitation amounts, as these can have large impacts on society. The Royal Netherlands Meteorological Institute (KNMI) is responsible for issuing weather warnings (for severe thunderstorms, among other weather hazards) and one criterion for a code yellow severe thunderstorm warning is the occurrence of local precipitation amounts exceeding $30\,\mathrm{mm\,h^{-1}}$. This threshold is greater than the 99th percentile in the afternoon period. A high-resolution, nonhydrostatic EPS has been running experimentally at KNMI (HarmonEPS; since autumn 2017), but it is too early to be used operationally. Further, it is expected that postprocessing of the output from this EPS is still necessary.

Model output statistics (MOS) is a regression-based statistical postprocessing technique that aims to correct deterministic forecasts for systematic errors in the dynamical model output by relating a response variable (such as observed precipitation) to one or more predictors from the numerical model, such as NWP model precipitation (Glahn and Lowry 1972). Traditionally, MOS used multiple linear regression to issue deterministic postprocessed forecasts that thus offered no information about forecast uncertainty, although later probabilistic forecasts were produced using logistic regression (LR; Lemcke and Kruizinga 1988). More recently, MOS has been extended to ensemble model output statistics (EMOS), which links the parameters of a forecast distribution to the characteristics of an EPS (such as the ensemble mean or spread). Here, we link the parameters of a forecast distribution to predictors from a deterministic model. Probabilistic forecasts for exceeding a single threshold can be issued from predictor variables using LR or for the whole forecast distribution using extended logistic regression (ELR; Wilks 2009). LR-based methods and EMOS (using various parametric distributions) have both been used extensively to postprocess many variables, such as temperature and sea level pressure (Gneiting et al. 2005), wind speed (Thorarinsdottir and Gneiting 2010; Baran and Lerch 2016), and lightning frequency (Schmeits et al. 2008). Statistical postprocessing of quantitative precipitation forecasts is challenging, compared to variables such as temperature or wind speed, given the positive probability of zero precipitation (point mass at zero in the forecast distribution), the high spatial heterogeneity of precipitation, and the increased uncertainty for high precipitation amounts (Scheuerer and Hamill 2015). Candidate distributions to calibrate precipitation forecasts are the generalized extreme value distribution (e.g., Scheuerer 2014), variants of the gamma distribution, including the censored and shifted gamma (CSG) and the zero-adjusted gamma (ZAGA) distributions (e.g., Sloughter et al. 2007; Bentzien and Friederichs 2012; Baran and Nemoda 2016; Scheuerer and Hamill 2015), and the lognormal or inverse Gaussian distributions (Bentzien and Friederichs 2012). These distributions are thought to be useful for modeling precipitation, given their heavy tails. Combining a logistic regression for the probability of precipitation with a gamma distribution has shown to be skillful, compared to other distributions (such as the lognormal or the inverse Gaussian) with less uncertainty (Bentzien and Friederichs 2012).

One limitation of ELR is its dependence on specific training thresholds. The thresholds chosen for training are critical in determining the most skillful parts of the forecast distribution. One limitation of a parametric approach is that a choice of distribution must be made. A nonparametric, data-driven approach may be able to avoid such assumptions (although surely making some others) while better managing nonlinear relationships. Examples of such nonparametric techniques are tree-based methods, such as classification and regression trees (CART; Breiman et al. 1984), random forests (RF; Breiman 2001), and quantile regression forests (QRF; Meinshausen 2006). Tree-based methods have been applied to many ''big data'' problems in recent years, including some implementations in climatology (Cannon et al. 2002; Whan et al. 2014), hydrology and streamflow forecasting (Chen et al. 2012; Galelli and Castelletti 2013), and statistical postprocessing (Carter and Elsner 1997; Gagne et al. 2014; Ahijevych et al. 2016; Taillardat et al. 2016; Gagne et al. 2017; Loridan et al. 2017; Taillardat et al. 2017; Herman and Schumacher 2018a,b).

Gagne et al. (2014) postprocessed threshold-based quantitative precipitation forecasts from a high-resolution mesoscale ensemble using random forests and multiple logistic regression. Forecast skill was comparable for methods that selected from an extended set of potential predictors (multiple logistic regression and random forests), with random forests displaying less resolution (as it was not able to forecast high probabilities). Extending the use of tree-based methods to a probabilistic framework, Taillardat et al. (2017) find that calibration of 6-h ensemble precipitation forecasts in France using quantile regression forests outperforms analog methods and compares favorably with a parametric approach (a censored and shifted gamma distribution).

Our work extends that of Taillardat et al. (2017) by including comparison of a tree-based method with additional traditional statistical postprocessing methods, using a large number of potential predictors from deterministic rather than ensemble NWP output, by focusing on very extreme hourly precipitation amounts and with the inclusion of additional verification measures. Quantile regression forests have also been used to generate probabilistic maps of wind speed thresholds associated with hurricanes (Loridan et al. 2017).

Here, we compare the skill and value of probabilistic precipitation forecasts issued by three statistical methods: 1) extended logistic regression, 2) a zero-adjusted gamma distribution, and 3) quantile regression forests, using predictors from a deterministic NWP model and with a focus on extreme precipitation. Description of datasets and methods is next (section 2), followed by results (section 3) and discussion and conclusions (section 4).

2. Data and methods

a. Datasets

The observed dataset is hourly calibrated radar precipitation in the Netherlands [calibrated against rain gauges; further information on the radar data and the calibration process can be found in Overeem et al. (2009)]. The high temporal (1-h accumulation) and spatial resolution (1 km^2) provided by the calibrated radar dataset allows postprocessing of precipitation on the local scale. The measured reflectivity in the radar signal on land is often influenced by tall objects that can result in spuriously high values that do not reflect real precipitation amounts (i.e., clutter). By looking at the annual rainfall sums of 2012, 2014, and 2016, a total of 43 radar cells were removed from the dataset (van Straaten et al. 2018). These land clutter points were likely caused by infrastructure in Rotterdam Harbor and The Hague and towers in Cabauw, Goes, and Hoogersmilde. However, no suitable correction was available for the occasional beam occulation in the dataset, and some clutter likely remains (Overeem et al. 2009). The calibrated radar dataset contains only the land points over the Netherlands, and we take $3 \times 3\,km^2$ averages. Average values, even over such small boxes, will be equal to or lower than any corresponding point measurements. We take such averages because we wish to avoid using a dataset where the most extreme precipitation values are based on only a single radar pixel. This method assumes that the heaviest precipitation events are associated with spatial extents much larger than 1 km^2 (Lochbihler et al. 2017), so the most extreme events will remain after averaging.

We divided the domain into 12 regions of approximately $80 \times 90\,km^2$, similarly to Schmeits et al. (2008), which allows the generation of area probabilistic forecasts. We excluded Region 1 from the analysis, as it has very few land points (Fig. 1). The response variable is the maximum $3 \times 3\,km^2$ radar precipitation in space (in each of the 11 regions) and time [the hourly maximum in each of four 6-h periods of a day: night (0000–0600 UTC; VT_0006), morning (0600–1200 UTC; VT_0612), afternoon (1200–1800 UTC; VT_1218), and evening (1800–2400 UTC; VT_1824)]. Taking area maxima from the regions and subsequent pooling allows us to increase the sample size, which is necessary for statistical forecasts of extreme events. The spatial pattern of the gridded hourly precipitation amounts and the maximum precipitation values for each region on a day with extreme precipitation are shown in Fig. 1. Figure 1 also shows the number of extreme precipitation ($Pr > 20\,mm\,h^{-1}$) in the afternoon period over the complete dataset (see below). Precipitation amounts are generally highest in the afternoon (1200–1800 UTC) and evening (1800–0000 UTC) periods (Table 1, Fig. 2). The observed precipitation quantiles for each verification time are shown in Table 2. In all analyses (including these tables and later figures), the values of all 11 regions have been pooled.

Potential predictor variables are taken from KNMI's high-resolution (2.5-km horizontal grid spacing), nonhydrostatic NWP model HARMONIE–AROME (HA). We use cycle 37h1.2, which assimilates both conventional observations (Bengtsson et al. 2017) and Mode-S aircraft data (de Haan 2011). We use a 3-yr reforecast dataset for the extended summer period (mid-April to mid-October) in 2010, 2011, and 2013. Reforecasts were initialized every 6 h (0000, 0600, 1200, and 1800 UTC) and run for +48 h. In addition to HA precipitation, we use 41 other potential predictors, including other direct model output (DMO) variables (at various pressure levels) and indices of atmospheric instability [Table 3; for further details on some indices, see Schmeits et al. (2005, 2008)]. Some of the most important indices of atmospheric instability are the Fateev index (based on differences between temperature at various pressure levels and the differences between temperature and dewpoint temperature), the modified Jefferson index (based on differences among wet-bulb potential temperature, temperature, and the temperature/dewpoint temperature difference, at specific levels), and the Boyden index (based on differences between geopotential height and temperature). Each of the potential predictors is treated

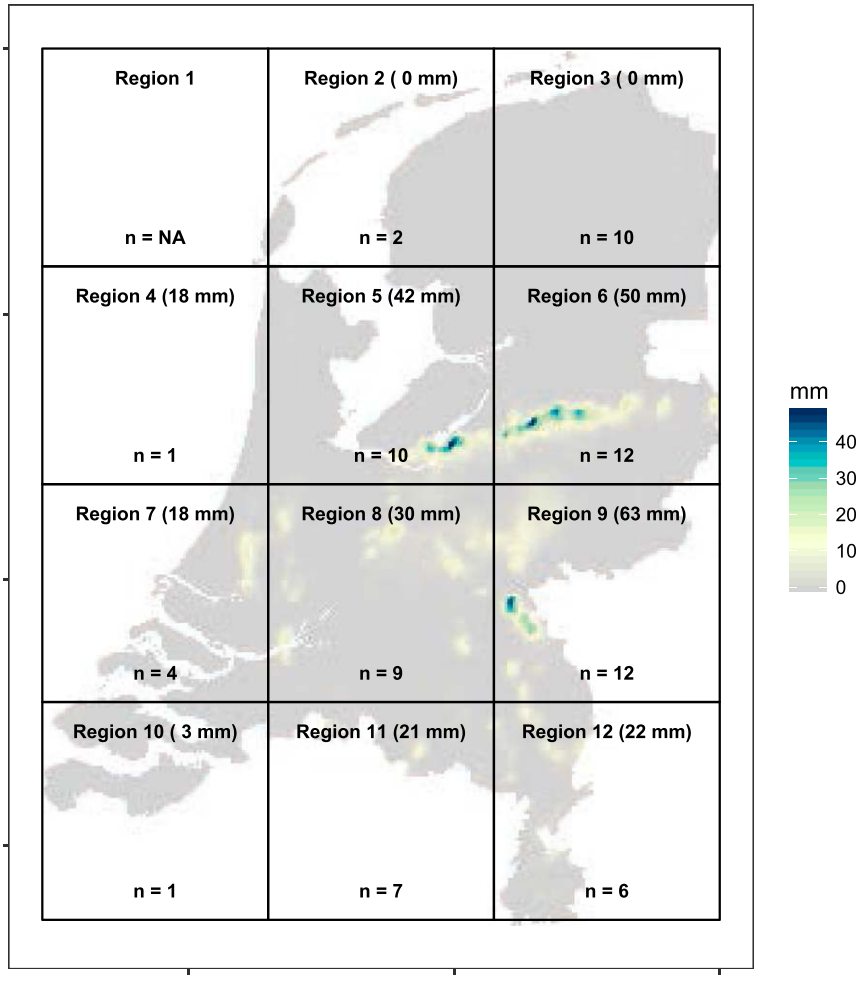

FIG. 1. The 12 regions ($80 \times 90\,\text{km}^2$) and an example of extreme hourly calibrated radar precipitation ending at 1700 UTC 28 Jul 2011. The maximum precipitation amounts in the 6-h afternoon period (1200–1800 UTC) are noted for each region in brackets following the region number. The counts (n; over all days) where precipitation exceeds 20 mm h^{-1} in the afternoon period are noted in the bottom of each region.

similarly to the response variable, as we take the maximum and minimum values in space (in each of the 11 regions) and time (maximum and minimum hourly values in each of the four 6-h periods). The region mask is defined on the radar grid. This mask is then bilinearly interpolated to the HA grid, and then the predictors are masked on their native model grid so that neither the radar nor model data is interpolated.

Improved model fit has previously been reported when a power transformation is applied to precipitation data [see Scheuerer (2014) and references within]. On a limited dataset, we test the skill of the forecasts using four combinations of transformed (cube root) precipitation as the response and predictor, where 1) both the predictor and response are transformed, 2) neither are transformed, and 3) only

TABLE 1. Counts of space–time maximum calibrated radar precipitation values in bins (mm h^{-1}) for each verification time.

	(0, 1]	(1, 5]	(5, 10]	(10, 20]	(20, 30]	(30, 50]	(50, 70]
VT_1218	4078	1082	497	234	51	20	3
VT_1800	4015	1010	437	192	45	11	1
VT_0006	4015	1081	327	128	16	1	0
VT_0612	4197	1174	393	164	34	5	1

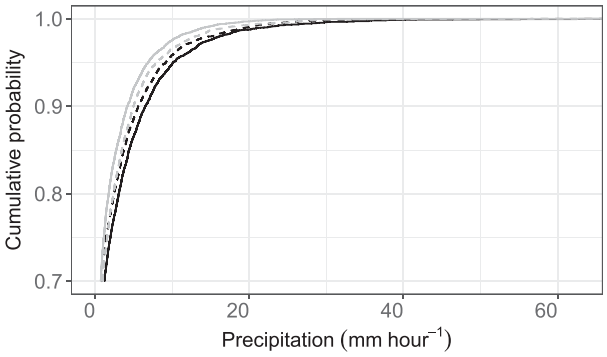

FIG. 2. Empirical cumulative distribution functions of calibrated radar precipitation over all regions for each valid time: afternoon (black, solid), evening (black, dashed), night (gray, solid), and morning (gray, dashed).

the predictor variable or 4) only the response variable is transformed. The highest skill is generally achieved for the parametric methods (ELR and ZAGA) when either both the response (observed precipitation) and predictor variables (HA precipitation) are transformed, or when only the predictor variable is transformed. The nonparametric method (QRF) shows fewer differences in forecast skill between the various combinations. As such, for future analyses, we only use transformed HA precipitation for the predictor.

b. Statistical postprocessing methods

We compare forecasts made using three statistical methods: ELR, ZAGA, and QRF. Given the relatively homogenous geography of the Netherlands, and to increase the number of extreme cases in the training dataset, values from all regions are pooled when fitting the statistical models. We use a threefold cross-validation framework to verify forecasts on an independent dataset comprising one summer half-year after training on the remaining two summer half-years. We then concatenate the three half-years of independent forecasts and verify them together against climatology from the full dataset. See appendix A for a description of the selection of optimum parameters in each statistical model.

We verify the probabilistic forecasts using the Brier skill score (BSS), reliability diagrams for various thresholds, and potential economic value (PEV; see appendix B for details on these metrics). We use block bootstrapping to calculate confidence intervals for the BSS. From a verification dataset of length n (three summer seasons), we draw 1000 random subsets of length n (with replacement) and take the cases for each of these n days from all 11 regions. We then calculate the BSS on each of these subsets using the climatology from the full record as a reference. We show the 95% confidence intervals of the BSS that are calculated from these 1000 bootstrap samples. We also computed confidence intervals for the reliability diagrams but do not present them here to maintain plot readability. The reader should keep in mind that for high precipitation thresholds, when the number of cases is small, the uncertainty around the reliability curves generally increases. We show sharpness diagrams for various precipitation thresholds in the afternoon period. These diagrams are representative of other verification periods.

TABLE 2. Observed quantiles of precipitation thresholds (0.3, 10, 20, 30 mm h^{-1}) in each verification period. Quantiles are determined by the rank in the dataset.

	0.3 mm h^{-1}	10 mm h^{-1}	20 mm h^{-1}	30 mm h^{-1}
VT_1218	0.583	0.950	0.988	0.996
VT_1800	0.607	0.960	0.991	0.998
VT_0006	0.604	0.976	0.997	1.000
VT_0612	0.587	0.966	0.993	0.999

1) EXTENDED LOGISTIC REGRESSION

First, we fit an ELR model. ELR is a simple logistic regression that includes a function of the threshold as a predictor and thus yields the full predictive distribution [Wilks 2009, Eq. (1)]:

$$p(q) = \frac{\exp[f(\mathbf{x}) + g(q)]}{1 + \exp[f(\mathbf{x}) + g(q)]}. \quad (1)$$

ELR predicts the probability of exceeding some threshold $[p(q)]$ from a linear function of predictors $[f(\mathbf{x})]$. The function $g(q)$ is a nondecreasing function of the threshold or quantile q, which allows the logistic regression equations for individual quantiles to be unified into one equation for any quantile (Wilks 2009). Here, $g(q) = a \times q$ is chosen. We use the generalized linear modeling function (glm) from the base package "stats" in the R statistical computing environment (R Core Team 2017). Predictors are chosen using forward and backward stepwise selection to minimize the Akaike information criterion (AIC; Akaike 1974; Sakamoto et al. 1986). The AIC estimates the fit of a statistical model based on the likelihood while penalizing more complex models [Eq. (2)]:

$$\mathrm{AIC} = -2[\log(L)] + 2K, \quad (2)$$

where L is the maximum likelihood function of the model, and K is the number of parameters. We limit the maximum number of predictors to avoid overfitting.

TABLE 3. Potential predictor variables, including DMO and indices of atmospheric instability. Definitions of the convection indices from temperature (T), dewpoint temperature (Td), Z (geopotential height), ff (wind speed), dd (wind direction), and θ (potential temperature) at various levels of the atmosphere (100, 500, 700, 850, 925, 1000 hPa), and c (the storm motion vector), u (zonal component of the wind), v (meridional component of the wind), k (vertical movement vector), ql (liquid water content), and g (acceleration due to gravity) are noted.

Predictor	Levels	Definition
θ_w	500, 850, 925 hPa	DMO: wet-bulb potential temperature (°)
θ_{ws}	500 hPa	DMO: wet-bulb pseudopotential temperature (°)
Boyden	Column	$0.1(Z_{700} - Z_{1000}) - T_{700} - 200$
Bradbury	Column	$\theta_{w500} - \theta_{w850}$
CAPE	surCAPE (surface), mulCAPE (most unstable layer)	$g\int_{LFC}^{LNB} \frac{T_v(\text{parcel}) - T_v(\text{environment})}{T_v(\text{environment})}\,dz$
Convective inhibition	Surface	$g\int_{\text{Surface}}^{LFC} \frac{T_v(\text{parcel}) - T_v(\text{environment})}{T_v(\text{environment})}\,dz$
Cross totals index	Column	$\text{Td}_{850} - T_{500}$
DPT	500, 600, 700, 850 hPa	DMO: Dewpoint temperature
Fateev	Column	$T_{850} - T_{500} - (T - \text{Td})_{850} - (T - \text{Td})_{700} - (T - \text{Td})_{600} - (T - \text{Td})_{500}$
HA precipitation	Surface	DMO: $\sqrt[3]{H - A}$ precipitation
Jefferson	Column	$1.6 \times \theta_{w925} - T_{500} - 11$
Jefferson (modified)	Column	$1.6 \times \theta_{w925} - T_{500} - 0.5(T - \text{Td})_{700} - 8$
K index	Column	$T_{850} - T_{500} + \text{Td}_{850} - (T - \text{Td})_{700}$
LFC	Column	Level of free convection
Lid strength	Column	$\theta_{ws_{\text{max\_lowest\_500hPa}}} - \theta_{ws_{\text{max\_lowest\_100hPa}}}$
Lifted index	Column	$\theta_{ws500} - \theta_{w100}$
LNB	Column	Level of neutral buoyancy
Precipitable water	Column	$\frac{1}{g}\int_{p2}^{p1} r\,dp$, where r = mixing ratio
Rackliff	Column	$\theta_{w925} - T_{500}$
Richardson (mulCAPE)	Column	$\text{mulCAPE}/(0.5 \times \text{Shear})^2$
Richardson (surCAPE)	Column	$\text{surCAPE}/(0.5 \times \text{Shear})^2$
S	Column	$\sin(\text{dd}_{500} - \text{dd}_{850})$
Shear	Column	$\sqrt{(u_{500} - u_{500\text{m}})^2 + (v_{500} - v_{500\text{m}})^2}$
Showalter	Column	$\theta_{ws500} - \theta_{w850}$
Storm-relative helicity	Column	$\int_0^{Z_{700}} k \times \left[(v - c) \times \frac{\delta v}{\delta z}\right] dz$
Storm travel	Column	$\max_\alpha\left\{\frac{\int_{\text{surface}}^{\text{highest level}} [u\cos(\alpha) + v\sin(\alpha)]\,dz}{\text{column height}}\right\}$, $0 \le \alpha \le \pi$
SWEAT	Column	Severe weather threat index: $12 \times \text{Td}_{850} + 20 \times (\text{Total totals index} - 49) + 2 \times 1.94\text{ff}_{850} + 1.94\text{ff}_{500} + 125 \times (S + 0.2)$
Total totals index	Column	$T_{850} + \text{Td}_{850} - 2 \times T_{500}$
TQ	Column	$(T_{850} + \text{Td}_{850}) - 1.7T_{700}$
UWND	700 hPa	DMO: eastward component of horizontal winds
Vertical totals index	Column	$T_{850} - T_{500}$
VWND	700 hPa	DMO: northward component of horizontal winds
WDIR	500, 850 hPa	DMO: the cosine of wind direction
WSPD	500, 850 hPa	DMO: wind speed (m s^{-1})

Further information on the parameter selection can be found in appendix A.

2) ZERO-ADJUSTED GAMMA DISTRIBUTION

Second, we fit a ZAGA distribution using the Generalized Additive Models for Location, Scale, and Shape (gamlss) package in R (Rigby and Stasinopoulos 2005; Stasinopoulos and Rigby 2007). The probability density function of ZAGA is defined by three parameters—the location μ, scale σ, and shape ν—and is defined as follows [Eq. (3)]:

$$f_Y(y|\mu,\sigma,\nu) = \begin{cases} \nu, & \text{if} \quad y = 0 \\ (1-\nu)\left[\dfrac{1}{(\sigma^2\mu)^{1/\sigma^2}} \dfrac{y^{(1/\sigma^2)-1}e^{-y/(\sigma^2\mu)}}{\Gamma(1/\sigma^2)}\right], & \text{if} \quad y > 0 \end{cases}, \tag{3}$$

with $\mu > 0$, $\sigma > 0$, and $0 \le \nu \le 1$, and where Γ is the gamma function. The ZAGA is a gamma distribution that allows mass at zero (i.e., the probability of zero precipitation), modeled by ν with a logit link function. The parameters of the ZAGA distribution are modeled as a linear function of the predictors. Maximum likelihood is used to fit the model.

3) QUANTILE REGRESSION FORESTS

Decision trees, such as CART (Breiman et al. 1984), seek to split a response variable into increasingly homogeneous ''nodes'' (groups) by minimizing a similarity measure [i.e., mean square error (MSE) in a regression framework] based on values of the predictors. They are easily interpretable but are known for easily becoming overfit and unstable, as small changes in the training dataset can result in very different trees and predictions (Hastie et al. 2001). A decision tree searches through all possible ''splits'' (break points in the predictors) and finds the split that results in the largest increase in node purity (largest decrease in MSE in the regression case). If left unchecked, a decision tree will eventually place each observation into ''terminal nodes'' (nodes with no further splitting) containing only a single value (i.e., MSE = 0). This overfitting can be reduced by ''early stopping.'' The RF method aims to minimize undesirable overfitting and decrease the variance of the forecast by building a set of m trees (e.g., $m = 500$) from m random subsets of the training dataset and using m random subsets of the potential predictors (Breiman 2001). In a regression setting, RF makes predictions of the conditional mean. Probabilistic forecasts are then possible using QRF (Meinshausen 2006). QRF extends RF by estimating the forecast cumulative distribution function. While RF only takes note of the mean of each node, QRF keeps all the values in the node and uses this to construct the conditional distribution (Meinshausen 2006). One benefit of nonparametric tree-based methods is that they do not assume that the response variable conforms to any particular distribution. This feature may be particularly important for variables like precipitation, where there is uncertainty about the distribution of the variable. We use the R package ''quantregForest'' (Meinshausen 2017), which builds upon the R package ''randomForests'' (Liaw and Wiener 2002).

The relative importance of potential predictor variables in a random forest is calculated by averaging the total decrease in node impurities when a predictor is used in the tree. In a regression setting, this is measured by the differences in errors before and after the split. Variables that reduce the errors more are then more ''important'' (Liaw and Wiener 2002). There is some randomness in the importance of the chosen predictors due to the high correlations between some of them. As such, we limit the discussion to the most important and consistent predictors.

3. Results

a. *Selected predictors*

Figure 3 shows the predictors that are selected in each of the cross-validation training sets for ELR and ZAGA (a predictor can be chosen in a certain position a maximum of three times) and the five most important predictors in QRF. HA precipitation is an important predictor that is selected by all methods at all lead times during the afternoon verification period (Fig. 3). It is selected as a predictor on two parameters of the ZAGA distribution, μ and ν. HA precipitation is the first predictor selected for the probability of zero precipitation (ν) at all lead times and is always followed by the Fateev index (Table 3). The two predictors chosen for μ are HA precipitation and the modified Jefferson index, with precipitation selected first at short lead times and second at longer lead times. Wind speed (maximum, 850 hPa) and thetaW (maximum and minimum, 850 and 925 hPa) are all selected as predictors for σ in one of the cross-validation training sets (Fig. 3). For all cross-validation sets and lead times, HA precipitation is the second predictor selected after the threshold by ELR. At short lead times, the Fateev index is selected third by ELR, while at longer lead times, the modified Jefferson index is selected third. The fourth predictor selected by ELR varies between cross-validation sets and lead times (Fig. 3). HA precipitation is the most important predictor in the QRF at all lead times. At short lead times, the next most important predictors are the Fateev index, CAPE (surface and the most unstable layer), and the lifted index. At longer lead times, the importance of the Fateev index decreases, and the importance of other indices, such as the level of neutral buoyancy (LNB), increases (Fig. 3). It is remarkable that CAPE (surface and the most unstable layer) is only selected

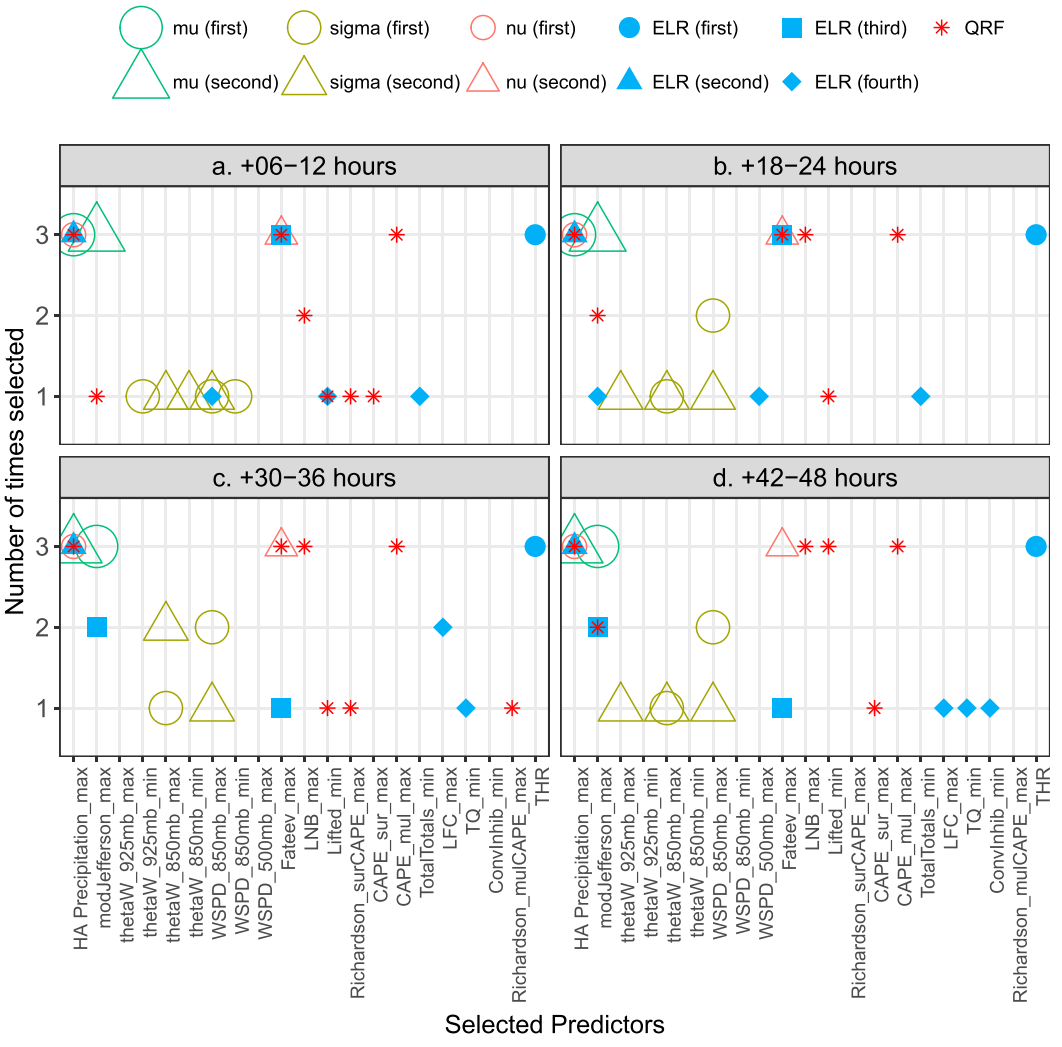

FIG. 3. Number of times in the cross validation (maximum = 3) at (a)–(d) different lead times during the afternoon verification period that selected indices are chosen as predictors for the parameters of the ZAGA distribution and the ELR model. The five most important predictors in QRF are also noted. THR is the threshold predictor in ELR.

by QRF, with the most unstable layer CAPE being selected more often than surface CAPE, as expected.

In the evening period, the importance of HA precipitation as a predictor for μ decreases with lead time, as it is the first predictor selected at short lead times and the second predictor at longer lead times, similarly to the afternoon period. However, contrary to the afternoon period, there is less consistency in the index of atmospheric instability that is chosen as the second predictor. At short lead times, the Bradbury, modified Jefferson, and Showalter indices are each selected second, each in one cross-validation fold, while at longer lead times, the modified Jefferson and LNB are selected as the first predictors (Fig. 4). Similarly, there is less consistency in the evening period for the predictors selected for ν, compared to the afternoon period. At the shortest lead time, the Boyden index is the second predictor, but at other lead times, the second predictor is divided between the Boyden and Fateev indices (Fig. 4). The signal-to-noise ratio can be relatively small, and the correlations between some predictors are relatively large. Random day-to-day differences between potential predictors likely have some influence on the final selected predictors, as evidenced by the differences in the predictors selected between verification periods and lead times. HA precipitation is an important predictor for QRF and ELR at all lead times. In the evening period, for ELR, the Fateev index is the third predictor (after the threshold and HA precipitation) at short lead times, while the

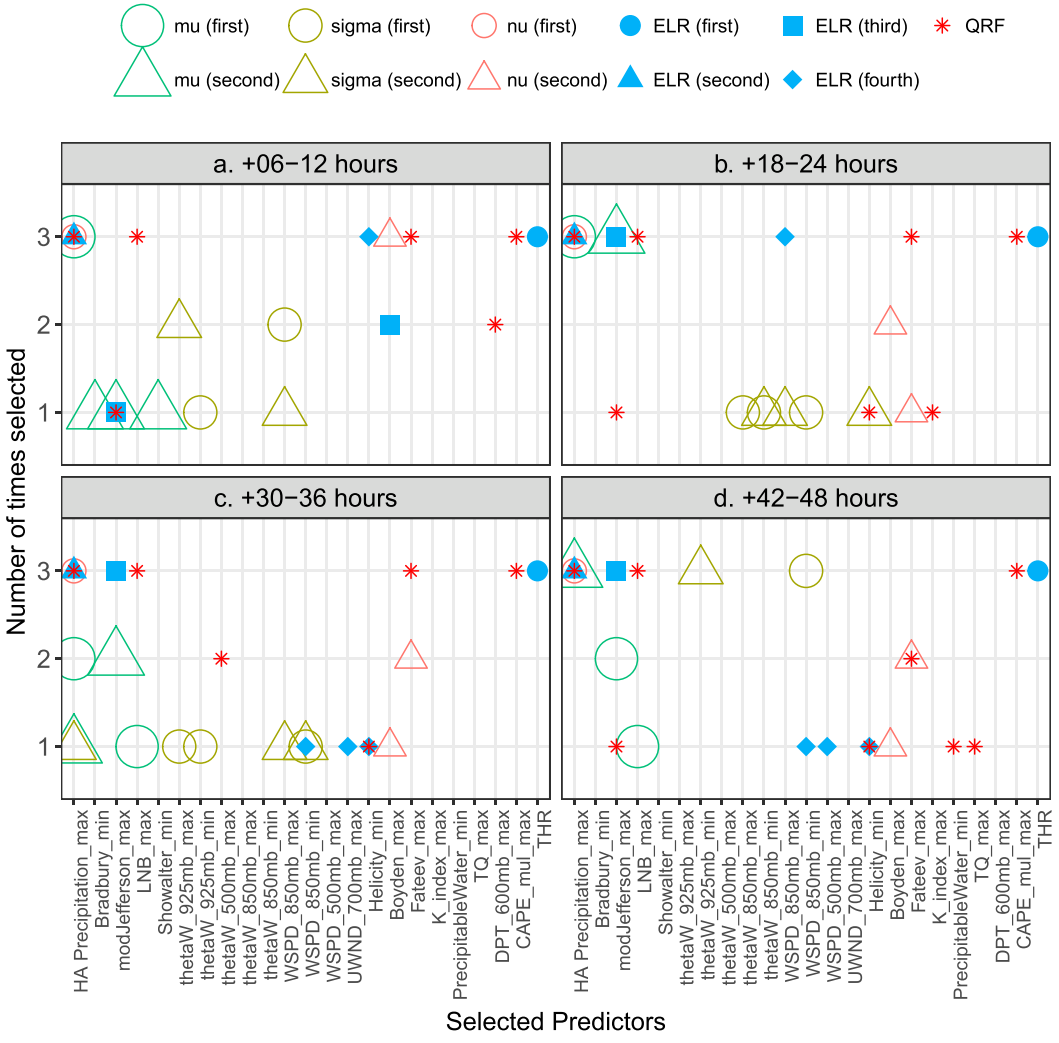

FIG. 4. As in Fig. 3, but for the evening verification period.

modified Jefferson index is the third predictor at longer lead times.

There are few differences in the correlation between observed and HA precipitation between the afternoon and evening (i.e., $r = 0.57$ to 0.59 at the shortest lead time), and the correlation decreases with increasing lead time (i.e., $r = 0.58$ to 0.5 for the afternoon verification period; Fig. 5a). The variable importance for HA precipitation in QRF decreases substantially with increasing lead time, particularly in the afternoon period (Fig. 5b). The correlation between observed precipitation and CAPE (the most unstable layer) is relatively strong (e.g., $r = 0.51$ at the shortest lead time). CAPE is not selected as a predictor by the parametric models, although it is one of the most important predictors used by QRF (at least one version of CAPE is usually in the top five most important predictors). The variable importance of CAPE from QRF is not high, compared to HA precipitation at short lead times; however, the relative importance increases with increasing lead time, as the importance of HA precipitation decreases. There is a strong correlation between HA precipitation and CAPE (most unstable layer; $r = 0.70$ in the afternoon period, at the shortest lead time). Taken together, these results indicate that CAPE adds relatively little information on top of HA precipitation for short lead times. This is possibly because we take space–time maxima, which lessens the penalization of spatial and temporal errors in HA precipitation. For other predictors (Fateev index and the modified Jefferson index), the differences in correlations between afternoon and evening are small, and the correlation decreases only slightly with increasing lead time (e.g., in the afternoon, the correlation at the shortest and longest lead times for the modified Jefferson index is $r = 0.40$ and 0.38, respectively). The variable importance of the Fateev

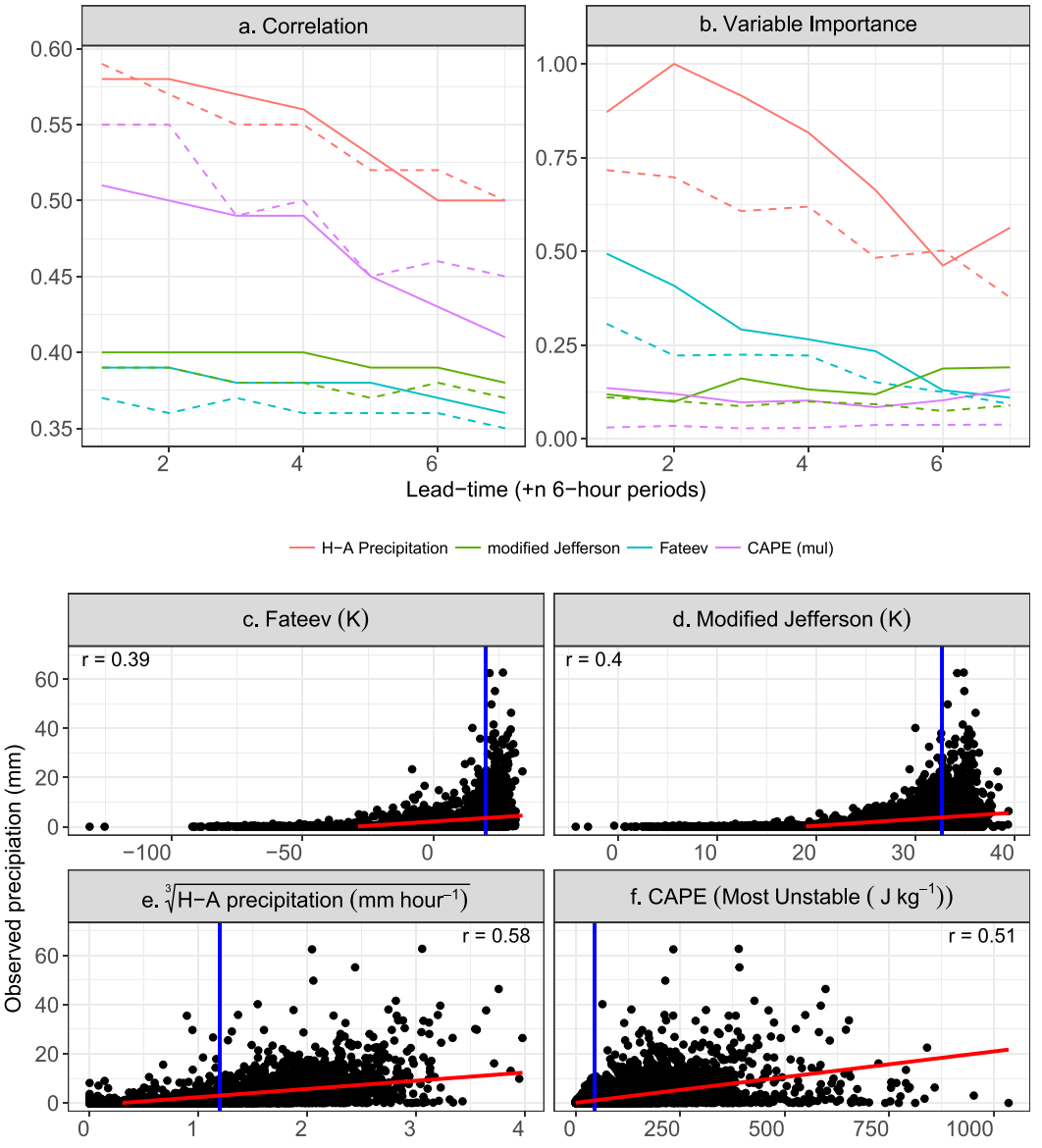

FIG. 5. (a) The correlation between a subset of potential predictors and observed precipitation and (b) the normalized variable importance for the same subset of predictors from QRF in the afternoon (solid) and evening (dashed) verification periods. The relationship between observed precipitation and selected potential predictors: (c) Fateev index, (d) modified Jefferson index, (e) HA precipitation, and (f) most unstable CAPE during the afternoon verification period and the +6–12-h lead time. Also shown are the correlation *r*, the linear regression (red line), and the most common split point selected for each potential predictor in the forest (vertical blue line).

index decreases with lead time, while the importance of the modified Jefferson index increases with lead time. There are substantial differences between verification times in the increases in node purity gained when splits are made from HA precipitation (Fig. 5b).

We compare the skill (using the BSS) of statistical models fit with HA precipitation only or the full set of potential predictors for each verification time and over four lead times (Figs. 6, 7). The value of allowing the statistical models to select predictors from the extended set of potential predictors (including other DMO and indices of atmospheric instability), in addition to the commonly used NWP model precipitation, is evident in the higher BSS for moderate thresholds for all methods and lead times (Fig. 6). Additional predictors are most important for QRF, as it has the largest difference in BSS between the rain-only model and the model using the extended set of potential predictors (Figs. 6, 7).

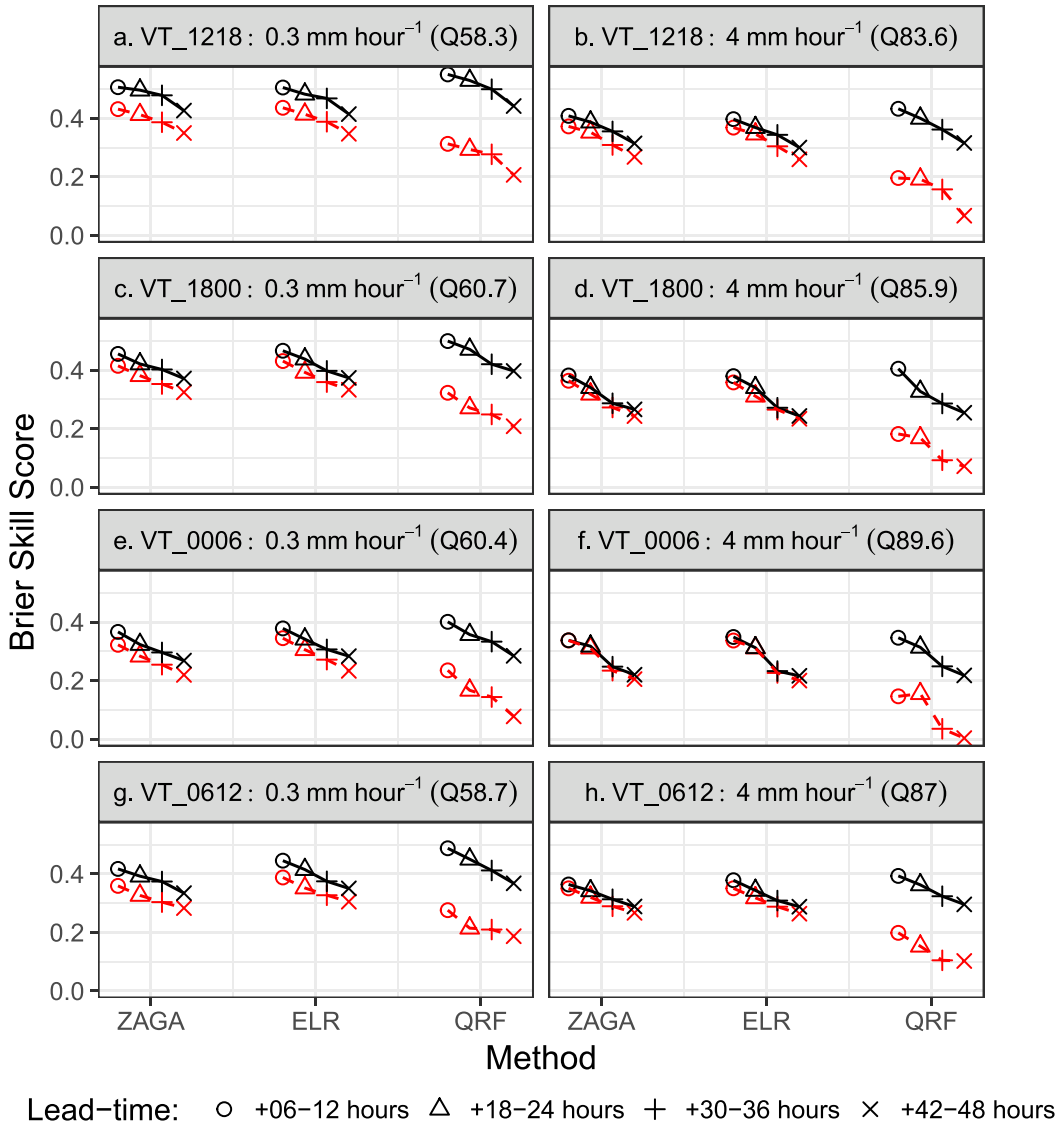

FIG. 6. The BSS for forecasts for low precipitation thresholds in all verification periods: (a),(b) VT_1218; (c),(d) VT_1800; (e),(f) VT_0006; and (g),(h) VT_0612 from models fit using three statistical methods (ZAGA, ELR, and QRF) and two sets of potential predictors: HA precipitation only (red; dashed) and including all DMO and indices of atmospheric instability as potential predictors (black, solid). Results are shown for four lead times (+6–12, +18–24, +30–36, and +42–48 h; indicated by the shapes) and two thresholds: (a),(c),(e),(g) 0.3 and (b),(d),(f),(h) 4 mm h^{-1}. The quantiles for each precipitation threshold in each verification period are indicated in the figure title.

Tree-based methods are not designed to be used with only a single predictor, but in such a situation, they make linear splits on the single predictor. Indeed, homogeneous terminal nodes could be created with a single predictor variable if the relationship between the predictor and the response was linear. The large differences between the QRF models using only HA precipitation and that using the full set of potential predictors shows the value of including additional predictors and may indicate possible nonlinearities in the relationship between the selected predictors and the response variable (Taillardat et al. 2016; Fig. 6). Indeed, scatterplots show that the relationships between some selected indices of atmospheric instability and observed precipitation are nonlinear (Figs. 5c–f). These nonlinear relationships do not prevent the selection of these indices in the linear, parametric framework, but it is likely that the relationships can be better captured by a nonlinear tree-based method that can also easily handle correlations between predictors.

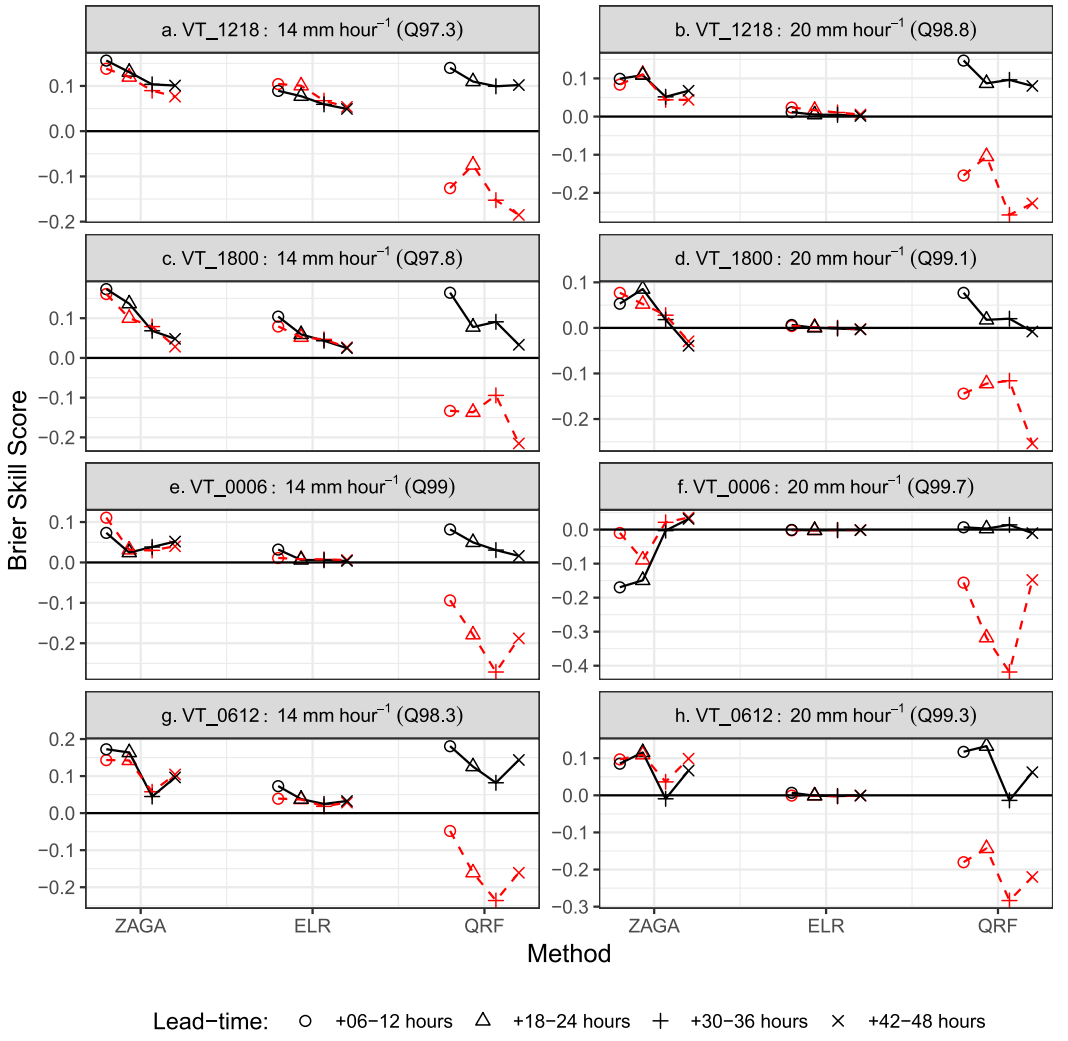

FIG. 7. As in Fig. 6, but for high-threshold precipitation forecasts (14 and 20 mm h^{-1}).

At higher thresholds, additional predictor information remains necessary for more skillful forecasts with QRF, but the benefits for forecasts using ZAGA are only evident during the afternoon and evening verification periods (Figs. 7a–d) and are even absent for some verification and lead times (e.g., short lead times during the night period; Figs. 7e,f). Indeed, ZAGA forecasts using only HA precipitation as a predictor outperformed the models using all potential predictor variables at longer lead times for higher precipitation thresholds during the morning verification period (Figs. 7g,h) and for higher precipitation thresholds during the night period (Figs. 7e,f).

b. *Verification results*

Next, we compare the skill of forecasts made by the three statistical postprocessing models that have selected predictors from the full set of potential predictors. Figures 8 and 9 show the BSS and confidence intervals for each statistical postprocessing method over precipitation threshold for each verification time and four selected lead times. The decrease in forecast skill over forecast lead time is moderate (Figs. 8, 9). Low to moderate thresholds (0.3–10 mm h^{-1}) are the most skillfully forecast, compared to climatology by all methods, in all verification periods and over all lead times (night and morning verification periods in Fig. 9; not shown for the afternoon and evening verification periods). The highest skillfully forecast precipitation threshold varies by verification period, in accordance with climatology: precipitation amounts of 15–30 mm h^{-1} are skillfully forecast in the afternoon period (1200–1800 UTC; Figs. 8a–d), while in the night period (0000–0600 UTC), skillful forecasts can only be made until 10–20 mm h^{-1} (Figs. 9a–d).

QRF is generally the most skillful statistical postprocessing method in terms of the BSS, although ZAGA is the most skillful at some verification times, lead times,

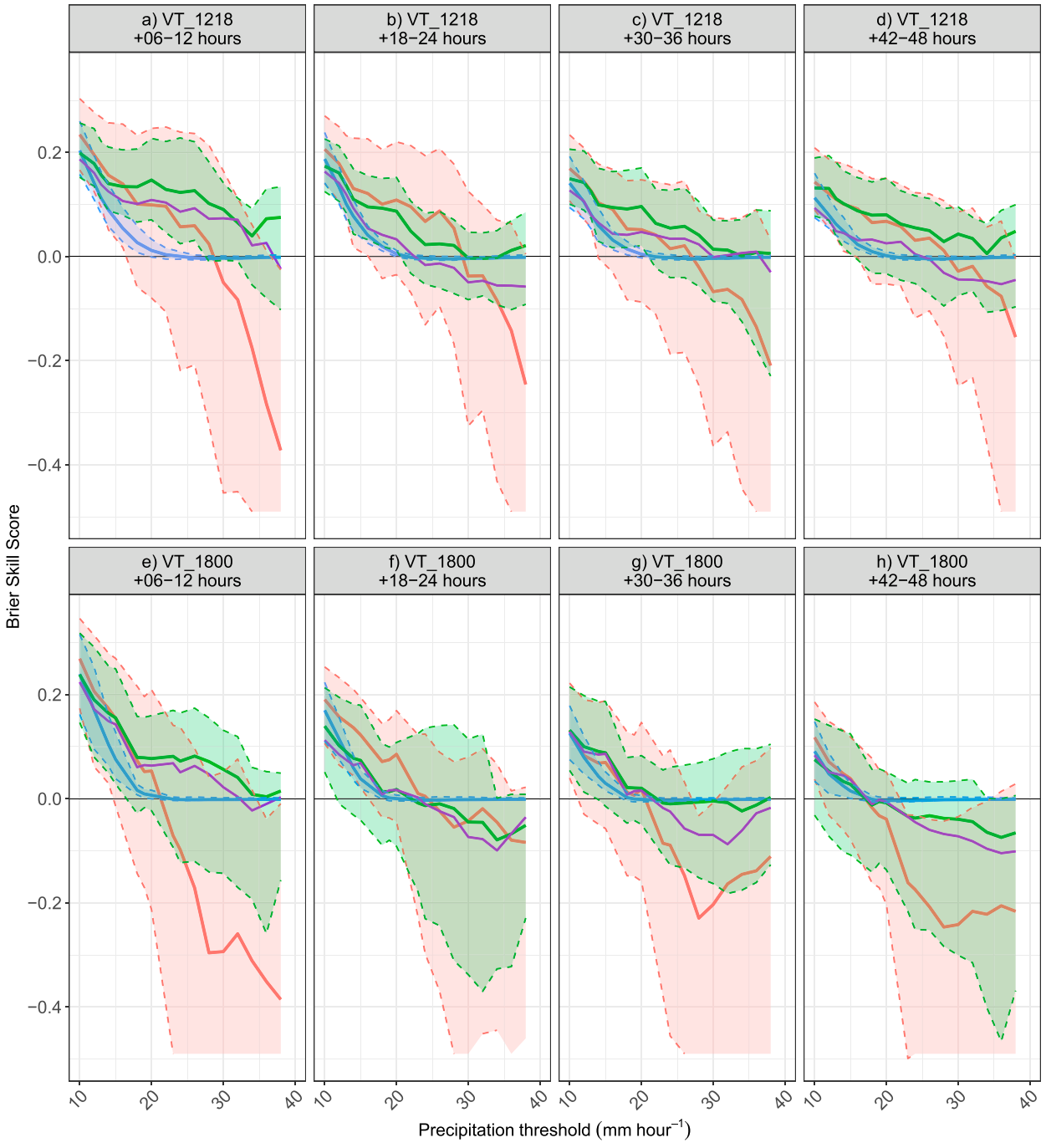

FIG. 8. The BSS for precipitation forecasts for thresholds between 10 and 40 mm h^{-1} in the afternoon (VT_1218) and evening (VT_1800) verification periods from models fit using three statistical methods [ZAGA: red, ELR: blue, QRF (all potential predictors): green, and QRF (DMO-only potential predictors): purple] at various lead times with HA precipitation and all atmospheric indices as potential predictors (see Table 3). Shading and dashed lines indicate the 95% confidence intervals of the BSS calculated from block bootstrapping.

and thresholds (e.g., Fig. 9d). The uncertainties in the QRF skill scores are also much lower, compared to ZAGA. For example, the skill of QRF and ZAGA is comparable for the 10–20 mm h^{-1} precipitation thresholds in the afternoon period at short lead times (Figs. 8, 9). However, the uncertainties for ZAGA are much larger and overlap the no-skill line, while for QRF, we can be confident that the method is skillful in this range (Fig. 8a). At higher thresholds, QRF outperforms ZAGA, and especially ELR, during most verification periods and forecast lead

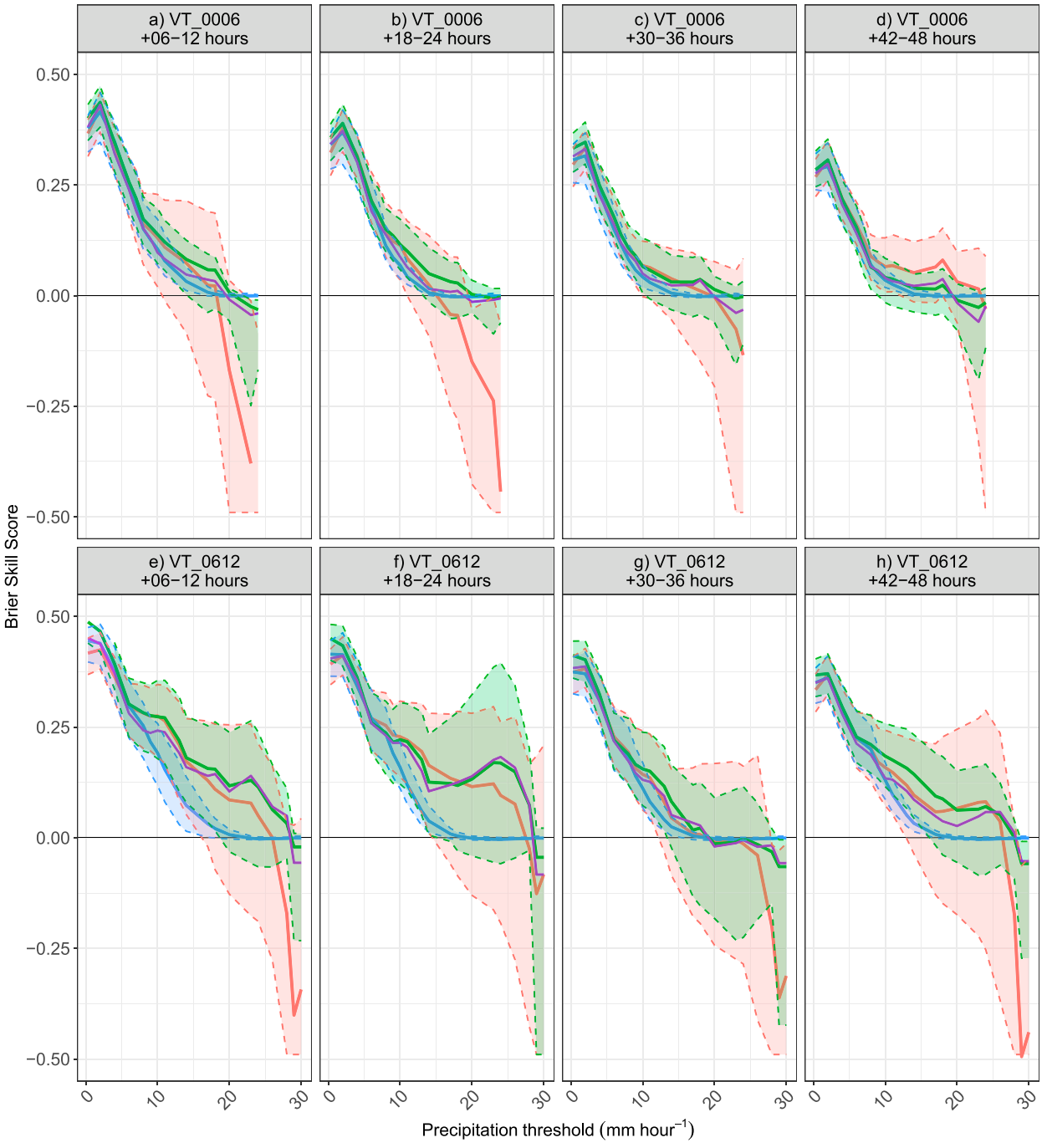

FIG. 9. As in Fig. 8, but for the night (VT_0006) and morning (VT_0612) verification periods and precipitation thresholds between 0.3 and 30 mm h^{-1}.

times, while at low and moderate thresholds, the comparative skill between methods is more mixed (Figs. 8, 9).

1) BSS: AFTERNOON PERIOD (1200–1800 UTC)

QRF is the most skillful method during the afternoon period (1200–1800 UTC; Figs. 8a–d), as it maintains skill with respect to the climatological probability (i.e., the confidence intervals of the BSS are above zero) for thresholds up to 20–30 mm h^{-1} at most lead times. During this period, the BSS uncertainty for high precipitation thresholds is large, but at the shortest lead time for QRF, the confidence intervals do not overlap

zero until the key threshold of 30 mm h^{-1} (Fig. 8a). ELR converges to zero skill at around 20 mm h^{-1}. For ZAGA, the BSS uncertainty at higher thresholds is much larger than for QRF, as it overlaps zero from around 15 mm h^{-1} (Figs. 8a–d).

Local precipitation amounts of 30 mm h^{-1} can be skillfully forecast in the afternoon period by QRF at the shortest lead time (+6–12 h). For the remaining verification and lead times, any postprocessed forecasts for this threshold have no BSS confidence intervals that are above zero, compared to climatology (Figs. 8, 9).

2) BSS: Evening, night, and morning periods (1800–0000, 0000–0600, 0600–1200 UTC)

In the evening period (1800–0000 UTC; Figs. 8e–h), ELR probabilities converge to the climatological probability around a threshold of 15–20 mm h^{-1}, above which ZAGA becomes less skillful than climatology at short lead times (Figs. 8e,f). The BSS confidence intervals for QRF overlap zero from around 20 mm h^{-1} at short lead times (Fig. 8e) and from around 10 mm h^{-1} at longer lead times (Figs. 8f–h).

Precipitation amounts in the night verification period (0000–0600 UTC; Figs. 9a–d) are lower and are thus predictable only at lower thresholds. QRF shows BSS confidence intervals above zero for thresholds up to 10–15 mm h^{-1}. At the longest lead time, ZAGA appears to be the most skillful method between 10 and 20 mm h^{-1}, although confidence intervals tend to overlap zero (Fig. 9d). Other methods converge to the climatological probability around this point (Figs. 9a–d).

In the morning verification time (0600–1200 UTC; Figs. 9e–h), ELR and ZAGA are not skillful anymore for thresholds exceeding 15 mm h^{-1} at short lead times, while QRF is more skillful than climatology until about 20 mm h^{-1} (Fig. 9e).

3) BSS: DMO-only predictors

The skill of QRF fit with only DMO as potential predictor variables is comparable to, or somewhat lower than, that of QRF fit with the full set of potential predictors (Figs. 8, 9). Where QRF is the most skillful method, the QRF model using only DMO predictors tends to also be more skillful than ZAGA and ELR; this is evident for high precipitation amounts in most verification periods at short lead times (Figs. 8a, 9e).

4) Reliability diagrams

Forecast probabilities for low precipitation thresholds are reliable using all postprocessing methods (Fig. 10). At higher thresholds, QRF does not issue forecast probabilities larger than 30%–50%, while ZAGA is overconfident and issues probabilities up to 90% (Figs. 10c,d,g,h). For moderate thresholds (e.g., 10 mm h^{-1}), all models tend to become somewhat overconfident at longer lead times while continuing to make a positive contribution to the BSS (Figs. 11c,d). Forecasts for higher thresholds (e.g., 20 mm h^{-1}), particularly with ZAGA, are overconfident from shortest lead time and do not make a positive contribution to the BSS at longer lead times (Figs. 11e–h).

5) Potential economic value

We show the PEV for the afternoon verification period at short lead times (+6–12 and +12–18 h) for the three statistical postprocessing methods and the deterministic HA precipitation forecasts (Fig. 12). PEV decreases with increasing precipitation threshold, but for a given threshold, the decreases in value between these lead times are moderate. In general, the differences between the statistically postprocessed forecasts from QRF and ZAGA are small, and probabilistic forecasts using these methods have more value than the deterministic forecast. QRF and ZAGA maintain value at high precipitation thresholds for low cost/loss (C/L) ratios (which makes up the majority of users; e.g., Figs. 12c,d), while ELR loses value. The statistically postprocessed forecasts using QRF and ZAGA generally have more value for a wider range of cost/loss ratios, compared to the raw HA precipitation forecasts.

4. Discussion and conclusions

Deterministic weather forecasts give no information about forecast uncertainty. Statistical postprocessing can generate probabilistic forecast information from ensemble or deterministic model output that yields this uncertainty. We compared the skill (defined by the Brier skill score) of probabilistic maximum local hourly accumulated precipitation forecasts in 11 regions of the Netherlands produced by three statistical methods (ELR, ZAGA, and QRF), using two sets of predictors from the deterministic HARMONIE–AROME (HA) model. This work demonstrates the production of probabilistic forecast information using predictors from deterministic NWP output, rather than the more commonly used ensemble forecast information. The sets of predictors were 1) only HA precipitation and 2) predictors selected from an extensive set of direct model output (precipitation, wind, and temperature) and indices of atmospheric instability. The inclusion of additional predictors (either in the stepwise selection

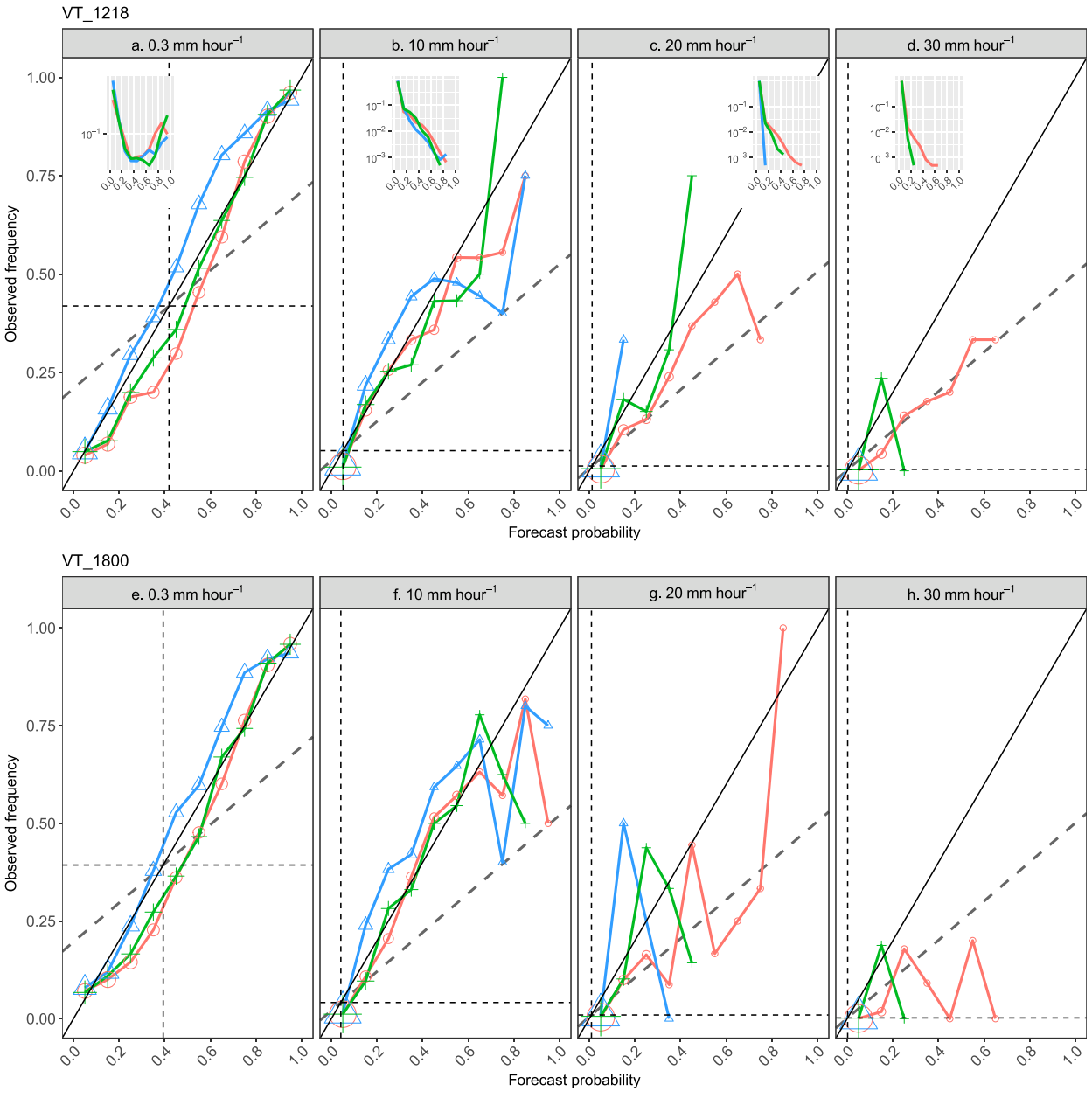

FIG. 10. Reliability diagrams for various precipitation thresholds (>0.3, 10, 20, and 30 mm h^{-1}) in the (a)–(d) afternoon and (e)–(h) evening periods for +6–12-h forecasts made using three methods (ZAGA: red, ELR: blue, and QRF: green) and using all the potential predictors listed in Table 3. As an indication, inset sharpness diagrams with a log scale are shown for each method in the afternoon period. Additionally, the size of the symbol indicates the number of forecasts in each bin.

step for the parametric methods or in QRF) results in a more skillful forecast for low and moderate thresholds in all verification periods and for extreme thresholds in all verification periods, but only for QRF (Taillardat et al. 2016). The inclusion of additional potential predictors results in larger increases in forecast skill, compared to using a different statistical postprocessing method. However, the value of including additional potential predictors is mixed for extreme precipitation thresholds for the parametric methods. The decrease in model skill from the shortest (+6–12 h) to the longest (+42–48 h) lead times is moderate. Further, we show the increased potential economic value of the probabilistic QRF and ZAGA forecasts, compared to the deterministic HA precipitation forecast.

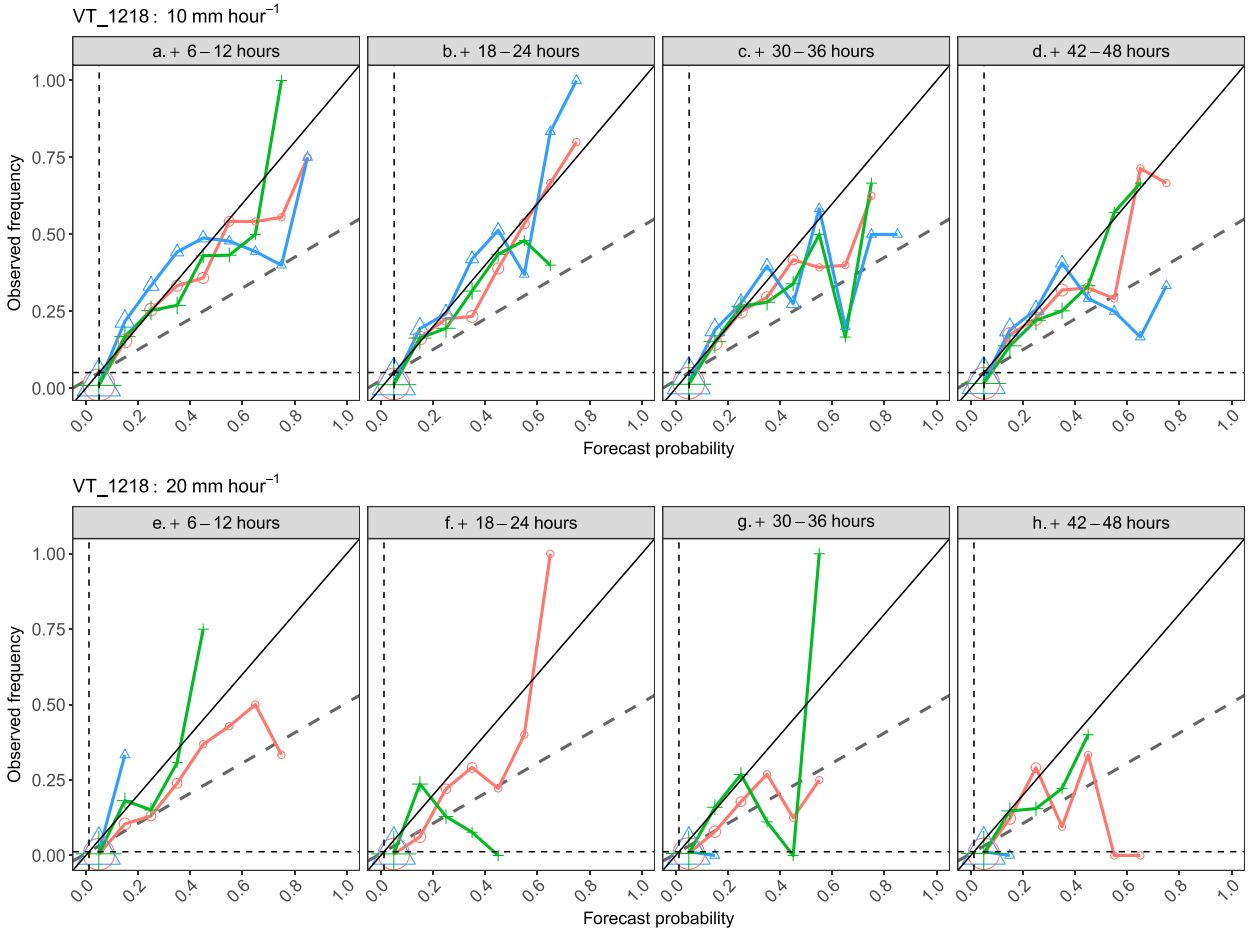

FIG. 11. Reliability diagrams for precipitation (a)–(d) >10 and (e)–(h) >20 mm h^{-1} during the afternoon verification period (VT_1218) and four lead times for models fit with the three methods (ZAGA: red, ELR: blue, and QRF: green) and using all the potential predictors listed in Table 3. The size of the symbol indicates the number of forecasts in each bin.

Quantile regression forests is the preferred method, as it is the most skillful with the least uncertainty in the BSS in the majority of lead times and precipitation thresholds, particularly for more extreme thresholds in the afternoon verification period. Indeed, the uncertainty in the BSS estimates from ZAGA were very wide, particularly for extreme thresholds, although we expect other parametric distributions to have higher uncertainties (Bentzien and Friederichs 2012). The high uncertainty in the BSS of the ZAGA distribution for large precipitation thresholds indicates that the forecast PDF is not well fit for extreme events. This is probably a result of the selected predictors and coefficients being heavily influenced by the bulk of the distribution. This may also suggest that the relationships between the response and predictors are different for low and high values of the response variable. This type of effect can be captured better by tree-based methods.

The nonlinear relationships between indices of atmospheric instability and observed precipitation are another possible reason why QRF outperforms the parametric methods. It is likely that further exploration and transformation of these potential predictors may result in more linear relationships and increased skill for the parametric methods, but such an exercise is outside the scope of the current study. One limitation of tree-based methods is that they cannot issue forecasts for precipitation amounts that are not in the training dataset, although in practice, this is not an issue for a multiyear training dataset, as there is no skill with respect to climatology for even the most extreme observed precipitation amounts using any method. A second limitation of QRF, which may have more practical implications, is that it does not issue higher forecast probabilities for extreme precipitation amounts (e.g., probabilities >60% are not produced for precipitation exceeding the 97th percentile in the afternoon period), similarly to Gagne et al. (2014).

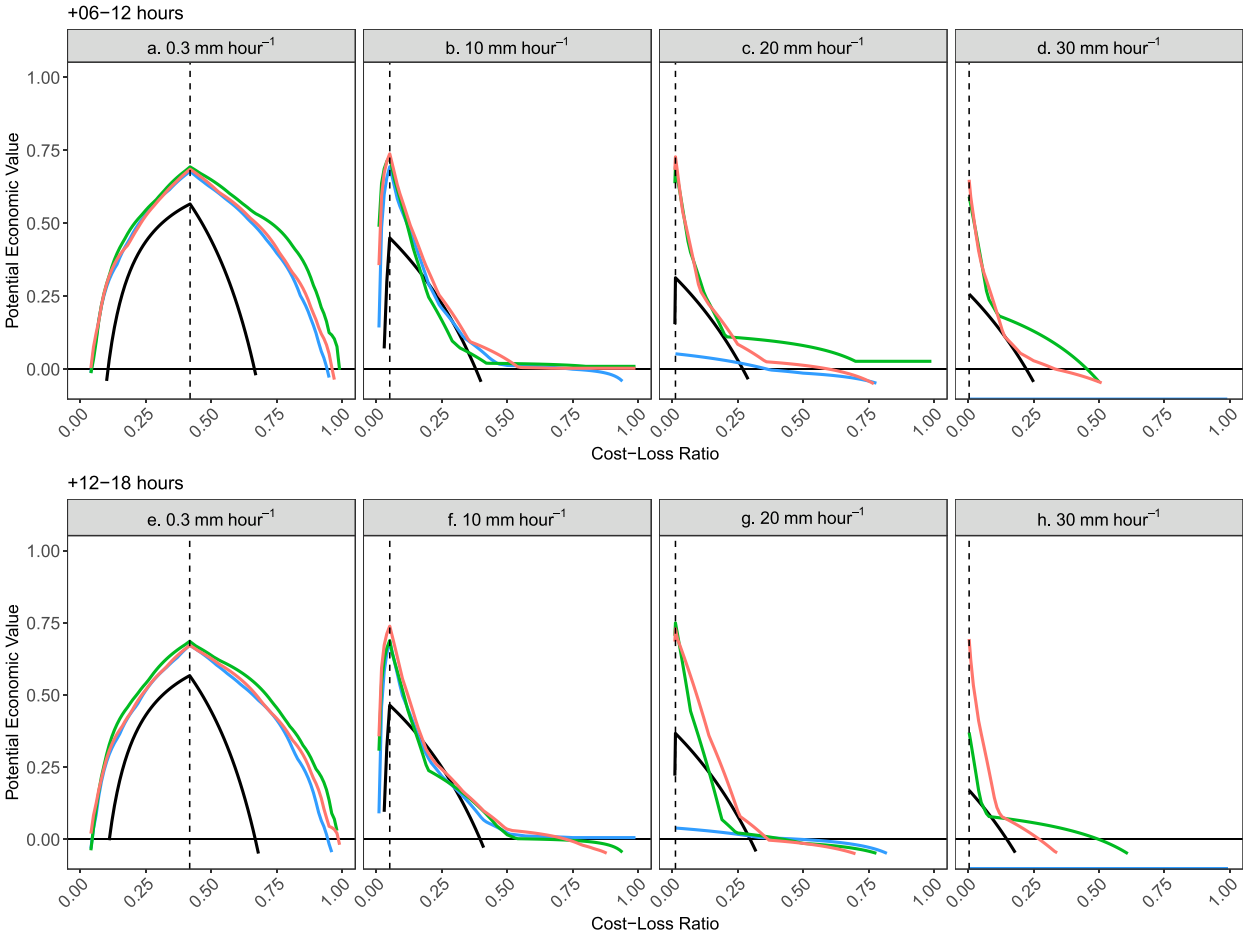

FIG. 12. Potential economic value as a function of C/L ratio for precipitation >0.3, 10, 20 and 30 mm h^{-1} during the afternoon verification period at (a)–(d) +6–12 and (e)–(h) +12–18 h lead times, for the deterministic HA precipitation forecast (black) and models fit with the three methods (ZAGA: red, ELR: blue, and QRF: green) and using all the potential predictors listed in Table 3.

Future work could compare other distributions, such as the GEV (Scheuerer 2014) or a formulation of ELR that does not depend on specific thresholds (Messner et al. 2014) with the nonparametric method or explore the influence of more elaborate methods to prevent overfitting, such as L1 regularization. Taillardat et al. (2017) found that QRF and a related method, gradient forests, compared favorably with parametric methods and outperformed an analog method. Future work could also explore possible increases in skill for the highest precipitation thresholds in this context that could be gained by extending the tail of the QRF distribution function using methods from extreme value theory, as in Taillardat et al. (2017). The inclusion of a more extensive set of potential predictors could result in a more skillful model [e.g., observations such as advected radar data could be included as an additional potential predictor source for short lead times, as in Schmeits et al. (2008)]. We have taken the maximum values of the response variable and predictors in the same regions, which partially decreases the influence of spatial displacement errors; however, in the future, this effect could be further reduced by including information from neighboring regions as potential predictors, as van der Plas et al. (2017) showed that spatial precipitation predictors from large neighborhoods are more skillful than those from smaller neighborhoods.

After a new reforecast dataset with the latest HA cycle has been generated, a preoperational probabilistic precipitation forecasting system will be developed using QRF to guide KNMI forecasters in the issuing of code yellow warnings for severe thunderstorms. Comparison of these statistical probabilistic forecasts with those issued directly by HarmonEPS will be explored in future work.

Acknowledgments. The authors wish to acknowledge several colleagues from KNMI. We thank Kees Kok for

useful discussions throughout this project. We thank Chiem van Straaten for help identifying clutter in the radar dataset. We thank Toon Moene for conducting the reforecasting experiment and Rudolf van Westrhenen, Erik van Meijgaard, and Cisco de Bruijn for programming part of the software to compute the indices of atmospheric instability. We also thank Gregory Herman and two other anonymous reviewers, whose comments improved the quality of this article.

APPENDIX A

Model Parameters

Several choices must be made for model parameters for each method. Ideally, such choices would be made on a separate independent dataset so as not to favorably bias the verification results. In a situation with limited data size, part of the dataset (e.g., 2 out of 3 years) can be used so that another part remains for verification. Given the focus on extremes and the limited reforecast dataset that is available, use of a truly independent dataset is not possible. However, the strong dependence of tree-based methods on the training dataset makes use of dependent data difficult, as these methods tend to give unrealistically skillful scores, resulting in unfair comparisons with the parametric methods. Therefore, we use threefold cross validation on the limited dataset. In this framework, the training dataset is 2/3 of 2 years, and the test dataset is the remaining 1/3 of the same 2 years (e.g., we train on 2/3 of 2010/11 and then test on 1/3 of 2010/11). In this way, the dataset we use to choose the optimum model parameters is partly independent from the dataset used to compare statistical postprocessing methods. The optimum parameter choices vary between verification times, and we base our partly subjective decisions primarily on the Brier scores (BSs) from the afternoon period when extreme rainfall amounts (and likely impacts) are largest. The decision to choose optimal model parameters based on the most important verification time means that skill in the other verification periods could likely be increased if models' parameters were chosen in a different way. An improvement to the current framework would be to choose model parameters that minimize the ranked probability score over a number of thresholds and per verification time and lead time.

a. ELR

We test the optimum values of the precipitation quantiles used to estimate the threshold predictor q (three sets of thresholds; Table A1) and the maximum number of selected predictors in addition to the threshold predictor (between one and eight). We fit models on the limited dataset with varying values of these parameters and calculate the BS.

TABLE A1. Possible values for the threshold predictor used in ELR.

Threshold group	Quantiles
Even	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9
Moderate	0.1, 0.5, 0.8, 0.85, 0.9, 0.925, 0.95, 0.99, 0.995, 0.999
High	0.5, 0.7, 0.8, 0.85, 0.9, 0.925, 0.95, 0.99, 0.995, 0.999

Based on these results (Fig. A1), we choose to use three predictors (in addition to the threshold) and the following quantiles for the threshold predictor: 0.5, 0.7, 0.8, 0.85, 0.9, 0.925, 0.95, 0.99, 0.995, and 0.999.

b. ZAGA

We compare forecast skill using a varying number of predictors on each parameter (between one and eight) and by using two stepwise selection methods that both select predictors based on the AIC. Method "A" fits a forward stepwise model to each parameter sequentially, given the selected model for the previous parameter (first μ, then σ, and finally ν). At each forward selection step, the full set of remaining predictors is tested. Next, backward stepwise selection is used in the reverse order to eliminate unnecessary predictors. In method "B," each of the potential predictors is fitted to all parameters at once [see Rigby and Stasinopoulos (2005) for further information]. Differences between stepwise selection methods are small for a low maximum number of predictors but increase for a higher maximum number of predictors. The highest skill is generally found for a low maximum number of predictors. According to these results, we choose to use default stepwise selection method "A" and two predictors per parameter (Fig. A1).

c. QRF

For QRF, we test the influence of the number of trees in the forest (between 100 and 1000) and the node size (an early stopping measure that limits the minimum number of values in a terminal node; between five and 80). Differences in skill according to the node size are larger than the differences between the numbers of trees. Based on these results, we choose to use the default value for both the number of trees (500 trees) and the terminal node size (five; Fig. A1).

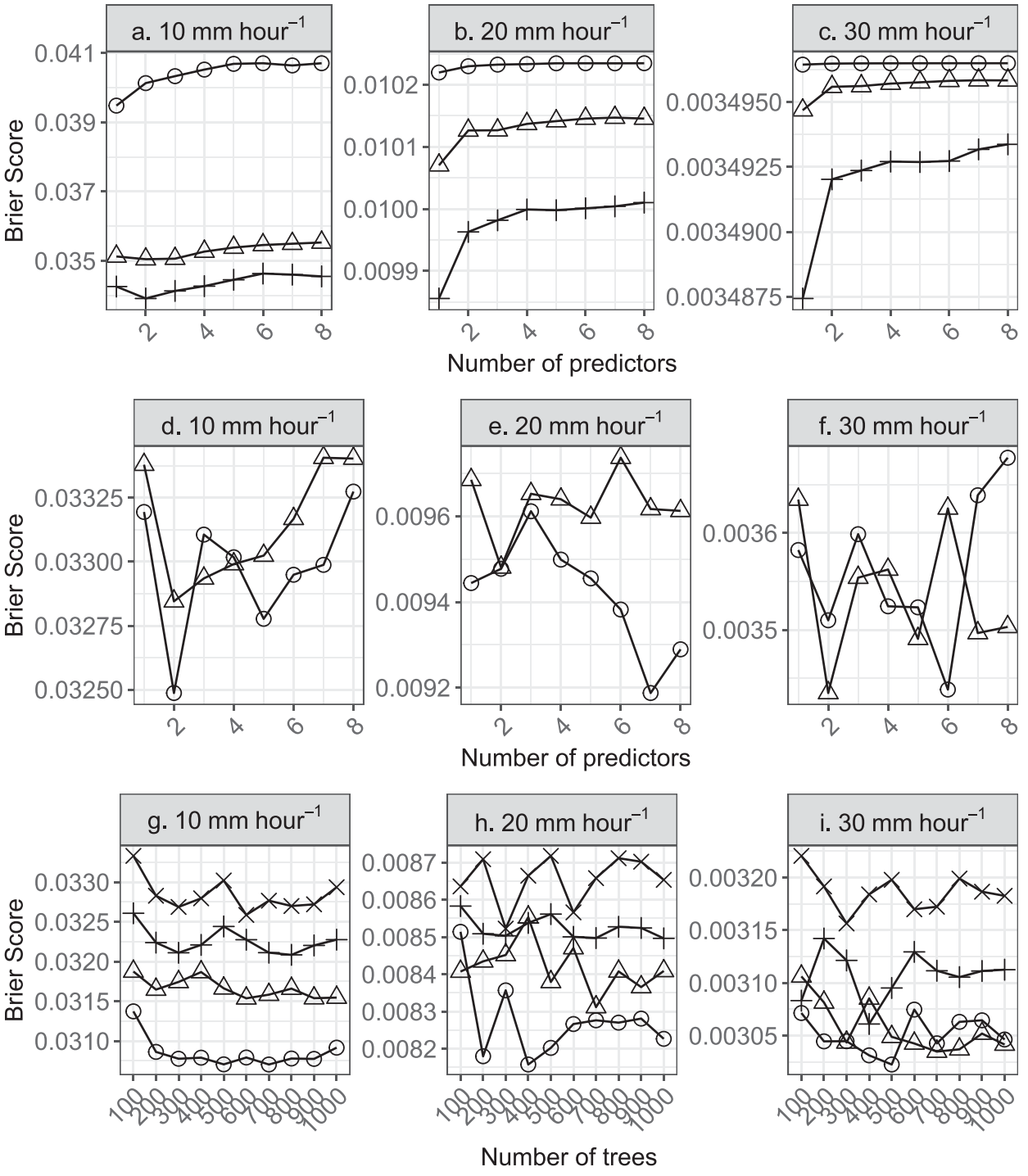

FIG. A1. BS for (a)–(c) ELR, (d)–(f) ZAGA, and (g)–(i) QRF models fit on a limited dataset (see text) with varying model choices in the afternoon period and the +6–12-h lead time. For ELR, we test skill using various quantiles as the threshold predictor (evenly spaced: circle, moderately high: triangle, high: plus; Table 4) and varying maximum number of predictors in addition to the threshold predictor. For ZAGA, we test model skill for a varying number of predictors and two stepwise selection methods (A: circle, B: triangle; see text for description of methods). For QRF, we test model skill for varying tree and node sizes (5: circle, 20: triangle, 40: plus, 80: cross).

APPENDIX B

Verification Metrics

a. Brier skill score

The BS is a measure of the magnitude of probability forecast errors (Brier 1950; Wilks 2011) so that a perfect score is zero. It is defined as the mean square error over all times $i = 1, 2, \ldots, N$ between the probabilistic forecast p_i and a binary observation o_i, which equals zero if the event did not occur and 1 if the event did occur [Eq. (B1)]. The negatively oriented BS is converted to the positively oriented BSS that shows the skill of the probability forecast relative to sample climatology [perfect score = 1; Eq. (B2)]:

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2, \tag{B1}$$

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{climatology}}}. \tag{B2}$$

b. Reliability diagrams

Reliability (or attribute) diagrams plot the observed frequency against the forecast probability over K bins (e.g., 0%–10%, 11%–20%, and 21%–30%). A reliable forecast is one where the forecast probabilities match the observed frequencies so that the line lies perfectly on the diagonal. For example, in a reliable forecast, an event that is predicted to occur 60% of the time is observed in 60% of the cases in which it is forecast. In an unreliable forecast, there is substantial mismatch between the forecast probability and the observed frequency. The size of the points indicates the number of data points in each bin. Regions that are halfway between the observed sample climatological frequency of an event and ''perfect reliability'' lines are colored gray. Forecasts in the gray area are said to be skillful, as they contribute positively to the BSS. Dashed lines indicate the climatological probability (Wilks 2011).

c. Potential economic value

PEV (Richardson 2000) can be plotted as a function of the C/L ratio. Probabilistic forecasts are converted to binary forecasts for all probability thresholds (Wilks 2011). The value score is then calculated from the hits (H; yes forecast/yes observed), false alarms (FA; yes forecast/no observed), misses (M; no forecast/yes observed), and climatological frequency $\overline{o}$, as in Eq. (B3):

$$\mathrm{PEV} = \begin{cases} \dfrac{(C/L)(H+FA)+M}{(C/L)(\overline{o}-1)}, & \text{if} \quad C/L < \overline{o} \\ \dfrac{(C/L)(H+FA)+M-\overline{o}}{\overline{o}[(C/L)-1]}, & \text{if} \quad C/L \geq \overline{o} \end{cases}. \tag{B3}$$

REFERENCES

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1.

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Automat. Control*, **19**, 716–723, https://doi.org/10.1109/TAC.1974.1100705.

Baran, S., and D. Nemoda, 2016: Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, **27**, 280–292, https://doi.org/10.1002/env.2391.

——, and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130, https://doi.org/10.1002/env.2380.

Bengtsson, L., and Coauthors, 2017: The HARMONIE–AROME model configuration in the ALADIN–HIRLAM NWP system. *Mon. Wea. Rev.*, **145**, 1919–1935, https://doi.org/10.1175/MWR-D-16-0417.1.

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, https://doi.org/10.1175/WAF-D-11-00101.1.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

——, J. Friedman, R. Olshen, and C. Stone, 1984: *Classification and Regression Trees*. Routledge, 368 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Cannon, A., P. Whitfield, and E. Lord, 2002: Synoptic map-pattern classification using recursive partitioning and principal component analysis. *Mon. Wea. Rev.*, **130**, 1187–1206, https://doi.org/10.1175/1520-0493(2002)130<1187:SMPCUR>2.0.CO;2.

Carter, M., and J. Elsner, 1997: A statistical method for forecasting rainfall over Puerto Rico. *Wea. Forecasting*, **12**, 515–525, https://doi.org/10.1175/1520-0434(1997)012<0515:ASMFFR>2.0.CO;2.

Chen, J., M. Li, and W. Wang, 2012: Statistical uncertainty estimation using random forests and its application to drought forecast. *Math. Probl. Eng.*, **2012**, 915053, https://doi.org/10.1155/2012/915053.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

de Haan, S., 2011: High-resolution wind and temperature observations from aircraft tracked by Mode-S air traffic control radar. *J. Geophys. Res.*, **116**, D10111, https://doi.org/10.1029/2010JD015264.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative

precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Galelli, S., and A. Castelletti, 2013: Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrol. Earth Syst. Sci.*, **17**, 2669–2684, https://doi.org/10.5194/hess-17-2669-2013.

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

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Hamill, T. M., 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.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning.* Springer, 536 pp.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

——, and ——, 2018a: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

——, and ——, 2018b: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Joslyn, S. L., and S. Savelli, 2010: Communicating forecast uncertainty: Public perception of weather forecast uncertainty. *Meteor. Appl.*, **17**, 180–195, https://doi.org/10.1002/met.190.

——, and J. E. LeClerc, 2012: Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *J. Exp. Psychol. Appl.*, **18**, 126–140, https://doi.org/10.1037/a0025185.

LeClerc, J. E., and S. L. Joslyn, 2015: The cry wolf effect and weather-related decision making. *Risk Anal.*, **35**, 385–395, https://doi.org/10.1111/risa.12336.

Lemcke, C., and S. Kruizinga, 1988: Model output statistics forecasts: Three years of operational experience in the Netherlands. *Mon. Wea. Rev.*, **116**, 1077–1090, https://doi.org/10.1175/1520-0493(1988)116<1077:MOSFTY>2.0.CO;2.

Liaw, A., and M. Wiener, 2002: Classification and regression by randomForest. *R News*, **2–3**, 18–22.

Lochbihler, K., G. Lenderink, and A. P. Siebesma, 2017: The spatial extent of rainfall events and its relation to precipitation scaling. *Geophys. Res. Lett.*, **44**, 8629–8636, https://doi.org/10.1002/2017GL074857.

Loridan, T., R. P. Crompton, and E. Dubossarsky, 2017: A machine learning approach to modeling tropical cyclone wind field uncertainty. *Mon. Wea. Rev.*, **145**, 3203–3221, https://doi.org/10.1175/MWR-D-16-0429.1.

Meinshausen, N., 2006: Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999.

——, 2017: quantregForest: Quantile Regression Forests. R package version 1.3-7, https://CRAN.R-project.org/package=quantregForest.

Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.

Morss, R. E., J. L. Demuth, and J. K. Lazo, 2008: Communicating uncertainty in weather forecasts: A survey of the U.S. public. *Wea. Forecasting*, **23**, 974–991, https://doi.org/10.1175/2008WAF2007088.1.

Overeem, A., I. Holleman, and A. Buishand, 2009: Derivation of a 10-year radar-based climatology of rainfall. *J. Appl. Meteor. Climatol.*, **48**, 1448–1463, https://doi.org/10.1175/2009JAMC1954.1.

Pinto, J. O., J. A. Grim, and M. Steiner, 2015: Assessment of the High-Resolution Rapid Refresh model's ability to predict mesoscale convective systems using object-based evaluation. *Wea. Forecasting*, **30**, 892–913, https://doi.org/10.1175/WAF-D-14-00118.1.

R Core Team, 2017: R: A language and environment for statistical computing. R Foundation for Statistical Computing, https://www.r-project.org/.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape (with discussion). *Appl. Stat.*, **54**, 507–554.

Sakamoto, Y., M. Ishiguro, and G. Kitagawa, 1986: *Akaike Information Criterion Statistics.* KTK Scientific Publishers, 290 pp.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, https://doi.org/10.1002/qj.2183.

——, 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.

Schmeits, M. J., K. J. Kok, and D. H. P. Vogelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting*, **20**, 134–148, https://doi.org/10.1175/WAF840.1.

——, ——, ——, and R. M. van Westrhenen, 2008: Probabilistic forecasts of (severe) thunderstorms for the purpose of issuing a weather alarm in the Netherlands. *Wea. Forecasting*, **23**, 1253–1267, https://doi.org/10.1175/2008WAF2007102.1.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, https://doi.org/10.1175/MWR3441.1.

Stasinopoulos, D. M., and R. A. Rigby, 2007: Generalized additive models for location scale and shape (GAMLSS) in R. *J. Stat. Software*, **23**, 1–46, https://doi.org/10.18637/jss.v023.i07.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

——, A.-L. Fougères, P. Naveau, and O. Mestre, 2017: Forest-based methods and ensemble model output statistics for rainfall ensemble forecasting. arXiv:1711.10937v1 [stat.ML], November 2017, https://arxiv.org/abs/1711.10937v1.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using

heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

van der Plas, E., M. Schmeits, N. Hooijman, and K. Kok, 2017: A comparative verification of high-resolution precipitation forecasts using model output statistics. *Mon. Wea. Rev.*, **145**, 4037–4054, https://doi.org/10.1175/MWR-D-16-0256.1.

van Straaten, C., K. Whan, and M. Schmeits, 2018: Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeor.*, https://doi.org/10.1175/JHM-D-18-0105.1, in press.

Whan, K., B. Timbal, and J. Lindesay, 2014: Linear and nonlinear statistical analysis of the impact of subtropical ridge intensity and position on south-east Australian rainfall. *Int. J. Climatol.*, **34**, 326–342, https://doi.org/10.1002/joc.3689.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, https://doi.org/10.1002/met.134.

——, 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.