Accounting for Skew when Postprocessing MOGREPS-UK Temperature Forecast Fields

SAM ALLEN,[a] GAVIN R. EVANS,[b] PIERS BUCHANAN,[b] AND FRANK KWASNIOK[a]

[a] *Department of Mathematics, University of Exeter, Exeter, United Kingdom*
[b] *Met Office, Exeter, United Kingdom*

(Manuscript received 21 December 2020, in final form 22 April 2021)

ABSTRACT: When statistically postprocessing temperature forecasts, it is almost always assumed that the future temperature follows a Gaussian distribution conditional on the output of an ensemble prediction system. Recent studies, however, have demonstrated that it can at times be beneficial to employ alternative parametric families when postprocessing temperature forecasts that are either asymmetric or heavier-tailed than the normal distribution. In this article, we compare choices of the parametric distribution used within the ensemble model output statistics (EMOS) framework to statistically postprocess 2-m temperature forecast fields generated by the Met Office's regional, convection-permitting ensemble prediction system, MOGREPS-UK Specifically, we study the normal, logistic, and skew-logistic distributions. A flexible alternative is also introduced that first applies a Yeo–Johnson transformation to the temperature forecasts prior to postprocessing, so that they more readily conform to the assumptions made by established postprocessing methods. It is found that accounting for the skewness of temperature when postprocessing can enhance the performance of the resulting forecast field, particularly during summer and winter and in mountainous regions.

KEYWORDS: Statistical techniques; Statistics; Ensembles; Probability forecasts/models/distribution; Statistical forecasting

1. Introduction

Surface temperature forecasts are of high demand to several industries, and also to the general public. It is therefore imperative that these forecasts are accurate and reliable, something that is typically not true for forecasts (either point forecasts or ensembles) generated from operational prediction systems. An a posteriori adjustment of the forecast is therefore necessary to alleviate systematic errors, while simultaneously quantifying the predictive uncertainty. To achieve this, state-of-the-art statistical postprocessing methods issue probabilistic forecasts in the form of predictive distributions. Although non- and semiparametric postprocessing approaches have recently received increased attention in the literature (e.g., Van Schaeybroeck and Vannitsem 2015; Taillardat et al. 2016; Henzi et al. 2020; Bremnes 2020), established postprocessing methods usually make distributional assumptions regarding the weather variable being forecast.

Several studies have therefore considered the appropriate statistical distributions to employ when postprocessing a range of weather variables. For nonnegative variables such as wind speed, the distribution should have a nonnegative support (Thorarinsdottir and Gneiting 2010; Messner et al. 2014; Scheuerer and Möller 2015), while that for precipitation should be nonnegative, but also contain a positive probability of being exactly zero (Sloughter et al. 2007; Scheuerer 2014; Scheuerer and Hamill 2015). For temperature, on the other hand, the normal distribution is almost invariably employed, both in a postprocessing context and throughout the wider field of atmospheric science (Von Storch and Zwiers 2001).

This is in part due to the appealing properties possessed by the normal distribution, which has led to its implementation in various branches of statistical modeling. For example, the normal distribution is widely used in linear regression (Klein et al. 1959; Glahn and Lowry 1972; Gneiting et al. 2005), time series models (Möller and Groß 2020), and spatial statistics (Scheuerer and König 2014; Scheuerer and Büermann 2014; Feldmann et al. 2015); it is conjugate to itself, and is thus convenient for Bayesian approaches (Stephenson et al. 2005; Siegert et al. 2016b; Barnes et al. 2019); while it also generalizes easily to multiple dimensions, making it the canonical choice for multivariate analysis (Schuhen et al. 2012; Feldmann et al. 2015; Barnes et al. 2019).

Because the normal distribution is so widely applied in studies concerning temperature, theoretical developments in statistical postprocessing models are often trialed first on temperature forecasts. The studies listed above are numerous examples of this, as are more recent approaches to ameliorate conventional postprocessing methods (Messner et al. 2017; Rasp and Lerch 2018; Schuhen et al. 2020). Nonetheless, Gebetsberger et al. (2018) have recently questioned the uninhibited use of the normal distribution as a means for postprocessing temperature forecasts. In particular, the authors suggest instead that the logistic and Student's t distributions are at times more appropriate, potentially because their heavier tails can account for the additional uncertainty that arises when estimating postprocessing parameters (Siegert et al. 2016a).

Moreover, the empirical distribution of temperature observations is regularly found to exhibit skew, often in particular seasons (Von Storch and Zwiers 2001). Although postprocessing is concerned with the conditional distribution of temperature given the numerical weather model output (and potentially other predictors) rather than its unconditional,

Allen's current affiliation: Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland.

Corresponding author: Sam Allen, sam.allen@stat.unibe.ch

DOI: 10.1175/MWR-D-20-0422.1

or climatological distribution, it is common to postprocess using a parametric distribution that resembles the climatological distribution of the response variable. In doing so, the forecast avoids assigning nonzero probabilities to weather events that cannot occur, while also capturing the limiting case where the outcome is independent of any predictors, in which instance the conditional distribution reverts to the variable's climatological distribution.

Therefore, Gebetsberger et al. (2019) propose recalibrating temperature forecasts using a Type-I generalized logistic distribution within the nonhomogeneous regression (NR), or ensemble model output statistics (EMOS), framework (Gneiting et al. 2005). The Type-I generalized logistic distribution extends the ordinary logistic distribution by including an additional shape parameter, thereby permitting skewed predictive distributions. Alternatively, asymmetric predictive distributions could be obtained by transforming the temperature forecasts prior to postprocessing, so that they conform to the assumptions made by more recognizable and convenient statistical methods, before applying the inverse transformation to the recalibrated forecasts. Hemri et al. (2015), for example, apply the well-known Box–Cox transformation to rainfall runoff before implementing nonhomogeneous Gaussian regression. There has previously been little interest in applying such transformations to temperature.

In this article, we consider both of these approaches to generate skewed predictive distributions when statistically postprocessing gridded temperature forecast fields over the United Kingdom, issued by the Met Office's high-resolution, convection-permitting ensemble prediction system, MOGREPS-UK MOGREPS-UK forecasts can be postprocessed using IMPROVER (https://github.com/metoppv/improver; Evans et al. 2020), a library of algorithms in development at the Met Office that utilize Rose and Cylc suites (Oliver et al. 2018, 2019) to postprocess and verify weather forecasts, and the work presented herein therefore builds upon the existing functionality within IMPROVER. In particular, we postprocess MOGREPS-UK temperature forecast fields using EMOS with a normal, logistic and Type-I generalized logistic distribution, and compare the resulting forecasts to those generated using nonhomogeneous Gaussian regression after having first applied a nonlinear transformation to the MOGREPS-UK ensemble output. It is demonstrated that relaxing the assumption of symmetry in the predictive distribution when postprocessing temperature ensemble forecasts can enhance the performance of the resulting forecast fields, particularly during summer and winter and in mountainous regions.

The model and data used to illustrate this are discussed in the following section. In section 3, we briefly discuss asymmetric variants of the normal and logistic distributions, as well as transformations that can be applied to address skew within samples of data. These approaches are then extended for use within the nonhomogeneous regression framework in section 4. Methods for parameter estimation and forecast verification are also discussed in section 4. Section 5 presents the performance of forecasts postprocessed using these variants of NR and comments on the choice of data to use when evaluating the performance of the gridded forecast fields. Finally, section 6 presents the conclusions.

2. Data

This study utilizes daily 2-m temperature forecasts extracted from the Met Office's MOGREPS-UK ensemble forecasting system (Hagelin et al. 2017) at lead times of 12, 24, and 36 h. The forecasts were issued in the 1-yr period between 1 January and 31 December 2018, during which time the model employed a 2.2-km horizontal resolution over the United Kingdom, and generated ensemble forecasts comprised of 12 members. The forecasts are initialized at 0300 UTC and thus validate at 1500 or 0300 UTC, allowing both day- and nighttime temperature predictions to be assessed.

Commonly, the aim of operational forecasting centers is to obtain a calibrated forecast field, from which predictions can be made for any location of interest. To do so, statistical postprocessing methods rely on an archive of historical forecasts and observations that adequately span the spatial domain under consideration, from which to learn previous errors of the prediction system. The spatial coverage afforded by weather recordings at synoptic stations, however, is typically inadequate: recordings over seas and oceans are generally particularly sparse (Hamill 2018). Postprocessing methods that utilize only the observations provided by this irregular network of stations are therefore unsuited to address the forecast biases at all locations on the domain, meaning systematic errors remain present in the postprocessed forecast field.

Instead, it would be desirable if the observations were available on a grid, similar to that of the forecasts. For this purpose, it is common to treat the analysis fields of a high-resolution numerical weather prediction model as the observations when postprocessing, rather than the recordings available at synoptic stations. The model analysis is the ''best guess'' of the atmospheric state at a particular time given the meteorological data to hand, as identified using data assimilation (Kalnay 2003). Although recent advances in data assimilation have made a significant contribution to the improved performance of numerical weather models (Alley et al. 2019), the analysis field is still prone to errors. Therefore, although using model analyses to train postprocessing methods accounts for data scarcity, the resulting forecasts typically underestimate the uncertainty present in reality (Feldmann et al. 2019).

An approach to combine both the model analyses and the weather observations at synoptic stations might be desirable when training postprocessing methods, but given the lack of such an approach, the postprocessing methods discussed herein are trained using model analyses. The analyses used are from the Met Office's deterministic, convective-scale UKV model, which operates on a domain with varying resolution, comprised of an inner domain with a resolution of 1.5 km, and a surrounding 4-km resolution area (Tang et al. 2013). Bilinear interpolation is used to map the MOGREPS-UK ensemble forecasts onto the smoother, inner domain of the UKV grid-space prior to postprocessing, and any further references to the UKV model domain relate to its inner domain. The result is a

forecast grid consisting of roughly half a million grid points (810 latitude points, 621 longitude points).

The UKV model domain is displayed in Fig. 1, along with the mean observed temperature field estimated over 2018. The average temperature generally decreases as latitude increases, with highest temperatures in the southeast of the United Kingdom and northern France, as expected. To demonstrate that the annual climatological temperature distribution is skewed, Fig. 2 displays the sample skewness of the temperature observations at each grid point, estimated over the same period. The sample skewness is defined as the Fisher–Pearson coefficient of skew, equal to

$$\frac{\sqrt{n(n-1)}}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{3/2}}, \tag{1}$$

where the temperature observations are denoted by y, with local mean $\bar{y}$, and n represents the number of observations from which the skewness is estimated. The empirical temperature distribution across the entire year is generally negatively skewed in the northwest region of the domain, whereas inland temperatures tend to be slightly positively skewed. The ensemble member forecasts tend to capture this general behavior well (not shown). Figure 2 also illustrates that the skew varies further in particular seasons, with the temperature more negatively skewed in winter and more positively skewed in summer. Since the negative skew in winter and the positive skew in summer are a result of the occurrence of more extreme low or high temperatures in these seasons, postprocessing methods that account for skew may be better suited to capture these more extreme weather events (Williams et al. 2014).

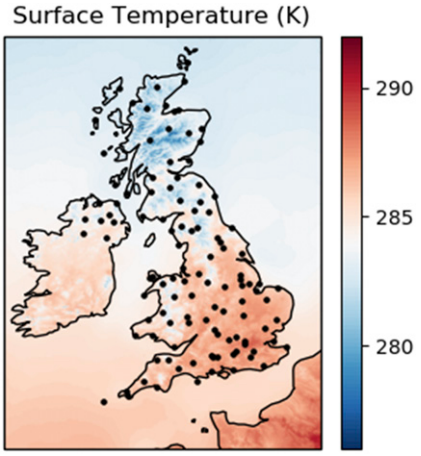

FIG. 1. Average observed temperature (K) across 2018 at 1500 UTC on the UKV model domain. The black points represent the 116 weather stations at which temperatures are considered in section 5b.

3. Accounting for skew

a. Skewed distributions

Although skewed distributions are not at all uncommon, recognized extensions of symmetric distributions, such as the normal and logistic distributions, to account for possible skewness are comparatively sparse. Azzalini (1985) introduced a very general class of skewed distributions whose probability density function (PDF) is of the following form:

$$g(y;\lambda) = 2f(y)F(\lambda y), \tag{2}$$

where f denotes a PDF that is symmetric about zero (e.g., the normal or logistic PDF), with corresponding cumulative distribution function (CDF) F. The shape parameter λ controls the skew of the distribution and, since f is symmetric, g encompasses f when this shape parameter is zero. Although distributions of this type are theoretically appealing, the resulting distribution functions are typically complex and difficult to manipulate (Gupta and Kundu 2010). Instead, the Type-I generalized logistic distribution (Johnson et al. 1995) provides a convenient, more accessible alternative that Gupta and Kundu (2010) argue is more appropriate for practical studies. As such, Gebetsberger et al. (2019) propose employing this distribution to postprocess temperature forecasts.

The probability density function of the Type-I generalized logistic distribution is

$$f_{\mathrm{GL}}(y;\mu,\sigma,\lambda) = \frac{\lambda\exp\left(-\frac{y-\mu}{\sigma}\right)}{\sigma\left[1+\exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{\lambda+1}}, \tag{3}$$

and its CDF is

$$F_{\mathrm{GL}}(y;\mu,\sigma,\lambda) = \frac{1}{\left[1+\exp\left(-\frac{y-\mu}{\sigma}\right)\right]^{\lambda}}. \tag{4}$$

The distribution is governed by a location parameter μ and positive scale σ and shape λ parameters. Unlike the skewed logistic distribution in the form of Eq. (2), the first four central moments of this generalized logistic distribution can all be expressed in closed-form, in terms of the polygamma function (Gupta and Kundu 2010). However, examples of this distribution's PDF in Fig. 3 suggest that it is unsuited to model heavily positive skew. Indeed, using the equation for the skew of the Type-I generalized logistic distribution presented in Gebetsberger et al. (2019) and properties of polygamma functions, it is possible to prove that the skewness of this distribution increases monotonically with the shape parameter λ, is bounded below by -2 and is bounded above by about 1.14. Yet further extensions of the logistic distribution exist—Johnson et al. (1995), for example, outlined four types of generalized logistic distributions, the simplest of which is the Type-I generalized logistic distribution characterized by Eqs. (3) and (4)–though, as with Eq. (2), increasing the complexity of the parametric distribution makes inference increasingly difficult.

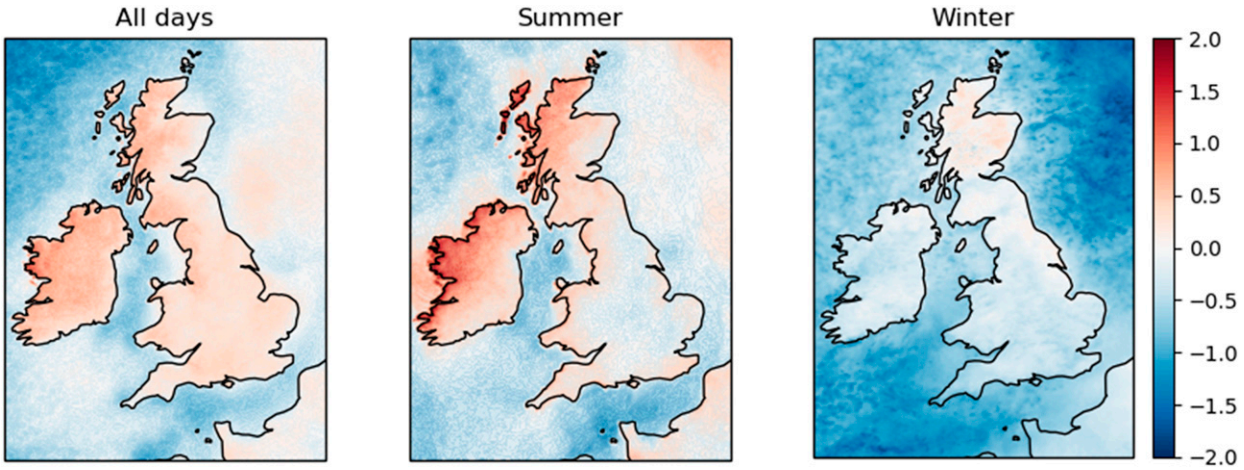

FIG. 2. Average sample skewness at 1500 UTC on the UKV model domain, shown for (left) all days, (center) all summer days, and (right) all winter days during 2018. The sample skewness is defined as the Fisher–Pearson coefficient of skew, as given in the text.

b. Transformations

Alternatively, rather than changing the distribution with which to represent temperature, transformations could be applied to temperature values so that they more readily conform to the assumptions made by particular, more desirable distributions. For example, it is often convenient to transform variables so that they appear more symmetric, allowing the implementation of more familiar statistical methods (Wilks 2019). As such, general purpose transformations have been developed to transform datasets so that they more closely resemble a sample from a Gaussian distribution. Arguably the most well-known example of this is the Box–Cox transformation (Box and Cox 1964).

However, the Box–Cox transformation is only suitable for nonnegative quantities. Although it is possible to include an additional shift parameter in the Box–Cox transformation that ensures all data are positive, Yeo and Johnson (2000) introduced a more unified approach to transform quantities defined on the entire real line:

$$\psi(z;\tau)=\begin{cases}[(z+1)^{\tau}-1]/\tau, & z\geq 0, \quad \tau\neq 0,\\ \log(z+1), & z\geq 0, \quad \tau=0,\\ -[(-z+1)^{2-\tau}-1]/(2-\tau), & z<0, \quad \tau\neq 2,\\ -\log(-z+1), & z<0, \quad \tau=2,\end{cases} \tag{5}$$

where τ is a parameter that controls the shape of the resulting distribution. When τ is equal to one, we recover the identity transformation. For values of τ smaller than one, on the other hand, the upper tail of the support is contracted, while the lower tail is extended, suggesting the variable at hand is positively skewed, whereas the opposite is true when $\tau > 1$. In the following sections, we compare this transformation with the Type-I generalized logistic distribution as a means of generating skewed predictive distributions when statistically postprocessing temperature forecasts. To maintain consistency with Gebetsberger et al. (2019), the Type-I generalized logistic distribution is hereafter referred to as the skew-logistic distribution.

4. Statistical postprocessing

a. Nonhomogeneous regression

The nonhomogeneous Gaussian regression (NGR) approach introduced by Gneiting et al. (2005) assumes that the future temperature is a random variable Y that follows a normal distribution with a mean that depends linearly on the mean of the ensemble member temperature forecasts $\overline{x}$ and a variance that depends linearly on their variance s^2:

$$Y|\mathbf{x} \sim N(\alpha_N + \beta_N\overline{x}, \gamma_N + \delta_N s^2), \tag{6}$$

where $\mathbf{x}$ denotes the vector of M ensemble members ($x_1, \ldots, x_M$), and α_N, β_N, γ_N, δ_N are parameters to be estimated. We discuss the nature of the parameter estimation in the following section. The two regression parameters for the distribution's location (α_N and β_N) address the biases in the ensemble mean forecast, while the two regression parameters controlling the

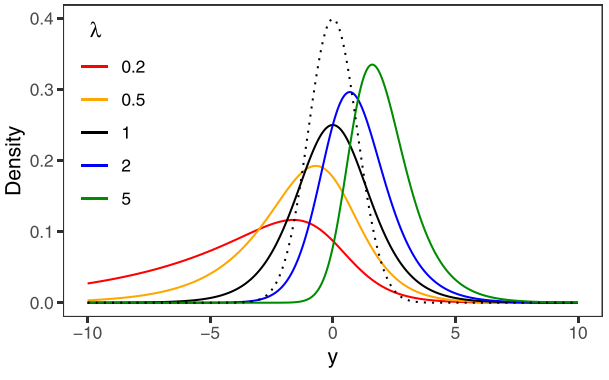

FIG. 3. Examples of the probability density function of the standard (i.e., location equal to zero, scale equal to one) Type-I generalized logistic distribution for various values of the shape parameter. The density function of the standard normal distribution is also displayed (dotted black line).

spread of the forecast distribution (γ_N and δ_N) account for dispersion errors in the ensemble.

This approach, which falls into the broad class of distributional regression methods (Klein et al. 2015), can be extended to employ alternative parametric distributions, in which case it is more generally referred to as nonhomogeneous regression (NR) or ensemble model output statistics (EMOS). Gebetsberger et al. (2018), for example, propose utilizing a logistic distribution within this framework:

$$Y|\mathbf{x} \sim L(\alpha_L + \beta_L \bar{x}, \sqrt{\gamma_L + \delta_L s^2}). \tag{7}$$

As in Eq. (6), the location parameter and the square of the scale parameter depend linearly on the ensemble mean and variance, respectively.

The skew-logistic distribution then extends Eq. (7) through the inclusion of an additional shape parameter:

$$Y|\mathbf{x} \sim L(\alpha_S + \beta_S \bar{x}, \sqrt{\gamma_S + \delta_S s^2}, \lambda_S), \tag{8}$$

where $L(\mu, \sigma, \lambda)$ represents the skew-logistic (i.e., Type-I generalized logistic) distribution with location μ, scale σ, and shape λ, whereas $L(\mu, \sigma)$ denotes the ordinary logistic distribution with location μ and scale σ (and shape equal to one).

Last, rather than changing the distribution to be used within NR, we consider a suitable transformation of the temperature forecasts and observations prior to postprocessing. Hemri et al. (2015) implement a similar approach whereby the Box–Cox transformation is applied to rainfall runoff before postprocessing the transformed forecasts using nonhomogeneous Gaussian regression. Since temperature is not constrained to be positive, we instead apply the Yeo–Johnson transformation [Eq. (5)] to the temperature data, but similarly implement nonhomogeneous Gaussian regression to postprocess the transformed forecasts. In particular, the approach proceeds as follows. The forecasts and observations in the training dataset are first standardized by removing the mean temperature observed in the training data, and dividing by the standard deviation of these temperature observations:

$$y_j^* = \frac{y_j - \bar{y}}{\sqrt{v_y}}, \quad x_{m,j}^* = \frac{x_{m,j} - \bar{y}}{\sqrt{v_y}}, \tag{9}$$

for all $m = 1, \ldots, M, j = 1, \ldots, N$, where j indexes over the N forecast–observation pairs in the training dataset, and $x_{m,j}$ represents the mth ensemble member on the jth forecast instance. The sample mean and variance of the N observations y_j in the training data are represented by $\bar{y}$ and v_y, respectively. Although this standardization could be performed individually for each grid point under consideration as a way to incorporate localized information (Dabernig et al. 2020), the mean and standard deviation are calculated across all grid points here to ensure a fair comparison with the alternative NR approaches considered in this study. Moreover, this standardization may not always be necessary, but it means the resulting temperature forecasts do not depend on the original unit of measurement.

Having standardized the forecasts and observations, the shape parameter of the Yeo–Johnson transformation is estimated by finding the value $\hat{\tau}$ that maximizes the (profile) likelihood of a Gaussian distribution given the transformed and standardized temperature observations in the training dataset, $\psi(y_j^*; \tau)$, as described in Yeo and Johnson (2000). The standardized temperature observations y_j^* and ensemble members $x_{m,j}^*$ in the training dataset are then transformed according to $\psi(\cdot\,; \hat{\tau})$, before fitting a nonhomogeneous Gaussian regression model [Eq. (6)] to these transformed, standardized forecasts and observations. Using the same value of τ to transform the temperature forecasts and observations ensures they remain on the same scale.

Unlike the variations of NR described above, this approach does not assume that the future temperature follows a particular parametric distribution, but rather that the standardized and Yeo–Johnson transformed temperature is normally distributed. Therefore, in order to generate a forecast for the future, untransformed temperature, it is necessary to sample from the predictive distribution issued by NGR for the transformed temperature, before applying the inverse of the Yeo–Johnson transformation, and finally recentering and rescaling using $\bar{y}$ and v_y.

b. Parameter estimation

The normal and logistic NR methods described above require four postprocessing coefficients to be estimated, whereas the skew-logistic distribution includes also a fifth. These parameters are generally estimated by minimizing a loss, or penalty function over a set of past forecasts and observations, referred to as the training dataset. The training dataset used in this study consists of forecasts issued during a rolling time window comprised only of the 30 days directly preceding the current forecast initialization time. In estimating a new set of coefficients for each forecast, this time-adaptive training window can account for the behavior of recent model errors (Gneiting et al. 2005).

To address locally varying biases when postprocessing, on the other hand, it is common to fit separate postprocessing models either at every location under consideration (Thorarinsdottir and Gneiting 2010), or for groups of locations based on proximity (Scheuerer and Hamill 2015), local climatological properties (Hamill et al. 2017; Friedli et al. 2021), or local characteristics of the forecast (Lerch and Baran 2017). Due to the extensive number of grid points considered here, elaborate methods to group together locations are computationally expensive, while site-specific postprocessing is infeasible. For this reason, one postprocessing model is fit to temperature forecasts across all grid points, which is the current framework implemented within IMPROVER. Such an approach has the benefit that the resulting postprocessing model can be applied to forecasts at all locations, including those for which past observations are not available. Moreover, because the postprocessing methods are applied to temperature forecasts aggregated over a substantially large number of grid points, there is always sufficient data from which to estimate reliable postprocessing coefficients: a 30-day rolling window consists of over 15 million temperature forecast–observation

pairs (30 × 810 × 621). Results presented in the following section are thus insensitive to the length of the training window (not shown).

There are two common choices for the loss function to use when estimating the coefficients over the training window. The first is the logarithmic or negative log-likelihood score, the minimization of which is equivalent to maximum likelihood estimation. Maximum likelihood estimation is used consistently throughout statistics due to its attractive theoretical properties (Gebetsberger et al. 2018; Wilks 2019) and it can easily be implemented for the skew-logistic forecast distribution using the probability density function given in Eq. (3). However, it has become routine in postprocessing studies to estimate parameters by minimizing the continuous ranked probability score (CRPS), since the resulting forecast distributions tend to be sharper (Gneiting et al. 2005). The CRPS is defined as

$$\mathrm{CRPS}(F,y)=\int_{-\infty}^{\infty}[F(u)-\mathbb{1}(u\geq y)]^2\,du, \tag{10}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Due to its continued use as a tool for both parameter estimation and also forecast verification, analytical solutions to this integral have been derived for several parametric distributions. Gneiting et al. (2005), for example, present a closed-form expression for the CRPS of a Gaussian predictive distribution:

$$\mathrm{CRPS}[N(\mu,\sigma^2),y]=\sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right)-1\right]+2\phi\left(\frac{y-\mu}{\sigma}\right)-\frac{1}{\sqrt{\pi}}\right\}, \tag{11}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the PDF and CDF, respectively, of the standard normal distribution, while Taillardat et al. (2016) and Jordan et al. (2017) derive the CRPS for the logistic distribution:

$$\mathrm{CRPS}[L(\mu,\sigma),y]=y-\mu-\sigma-2\sigma\log F_L\left(\frac{y-\mu}{\sigma}\right), \tag{12}$$

where F_L is the CDF of the standard logistic distribution $L(0,1)$.

To evaluate the CRPS for the skew-logistic distribution, Gebetsberger et al. (2019) use numerical integration techniques. In the appendix, we show that the CRPS for the standard skew-logistic distribution can be expressed as the following infinite series:

$$\mathrm{CRPS}[L(0,1,\lambda),y]=-\log F_L(y)+\sum_{k=1}^{\infty}\frac{1}{k}[1-F_L^k(y)]-2\sum_{k=0}^{\infty}\frac{1}{k+\lambda}[1-F_L^{k+\lambda}(y)]+\sum_{k=0}^{\infty}\frac{1}{k+2\lambda}. \tag{13}$$

The convergence of this series is fast, but becomes slower if there exist observations in the training data that lie in the extreme upper tail of the forecast distribution, that is, observations for which $F_L(y)$ approaches the radius of convergence, 1.

It is also shown in the appendix that there is an analytical representation of the CRPS for all rational values of the shape parameter (i.e., $\lambda = a/b$ with $a, b \in \mathbb{N}$). For example, if $a = 1$ and $b = 2$, then the CRPS becomes

$$\mathrm{CRPS}[L(0,1,1/2),y]=y+4\log\frac{1+F_L^{-1/2}(y)}{2}. \tag{14}$$

Similarly to the CRPS of the logistic distribution, we have that

$$\mathrm{CRPS}[L(\mu,\sigma,\lambda),y]=\sigma\mathrm{CRPS}\left[L(0,1,\lambda),\frac{y-\mu}{\sigma}\right], \tag{15}$$

and Eqs. (13) and (14) therefore easily extend to all possible values of the location and scale parameters.

Gebetsberger et al. (2018) argue that, since both estimators are consistent, maximum likelihood and minimum CRPS estimation should both yield calibrated forecasts if a suitable parametric distribution is employed in the statistical postprocessing model. However, if invalid distributional assumptions are made by the statistical model, then training the approach by minimizing the CRPS results in overly sharp forecasts, whereas the logarithmic score, in penalizing poorer forecasts more heavily, encourages the postprocessing method to overestimate the forecast spread. Therefore, to assess the distributional assumptions made by the different postprocessing methods compared in this study, parameter estimation is performed using both minimum CRPS and maximum likelihood estimation.

In all cases, to ensure the regression coefficients for the scale of the predictive distributions are positive, the loss function is minimized with respect to $\xi=\sqrt{\gamma}$ and $\kappa=\sqrt{\delta}$ rather than γ and δ directly, and the shape parameter of the skew-logistic distribution is similarly estimated using a square root link function to ensure positiveness. This shape parameter is estimated simultaneously to the model's other regression parameters, while the coefficients of the NGR postprocessing method applied to Yeo–Johnson transformed temperatures are estimated after having obtained $\hat{\tau}$, as described previously. However, due to the extensive amount of data provided by the high-resolution MOGREPS-UK forecast fields, minimum CRPS estimation for the skew-logistic distribution becomes computationally infeasible here, and results are therefore only presented using maximum likelihood estimation for this approach. Nonetheless, we demonstrate in section 5b that minimum CRPS estimation with the skew-logistic distribution is readily applicable to other settings where smaller training datasets are in place.

c. *Forecast verification*

Although nonhomogeneous regression issues forecasts in the form of predictive distributions, it is often more practical, and thus more common, to deal with a finite number of ensemble members. Therefore, after having postprocessed the raw forecast using nonhomogeneous regression, an ensemble is generated from the 12 evenly spaced (1/13, 2/13, . . . , 12/13) quantiles of the postprocessed forecast distribution. In the case of the transformed temperatures, these sampled quantiles are then subjected to the inverse Yeo–Johnson transformation to generate forecasts for the (untransformed) temperature, as is described at the end of section 4a.

Methods to verify ensemble forecasts can then be applied. The most common tool to assess ensemble forecasts is the rank histogram, which counts the frequency with which the observed temperature value assumes each rank when pooled among the ensemble members (Thorarinsdottir and Schuhen 2018). Deviation from uniformity in the rank histogram indicates a miscalibrated ensemble forecast, and systematic structures to this deviation can be used to diagnose the nature of any deficiencies in the prediction system (Hamill 2001).

Furthermore, the goal of probabilistic forecasting is often stated as increasing the sharpness of the forecast, subject to calibration (Gneiting et al. 2007). The coverage of the ensemble forecast is the proportion of instances in which the observed temperature falls between the lowest and highest ensemble members–more formally, this is the coverage of the forecast's $100(M-1)/(M+1)\%$ prediction interval; in our study, with $M = 12$, this corresponds to the forecast's 85% prediction interval. If the observation is equally likely to assume any rank among the ensemble members, then this coverage should be equal to $(M-1)/(M+1)$, the proportion of ranks the observation can take on while remaining between the lowest and highest ensemble member. The range, or width, of the ensemble members then provides a measure of the forecast sharpness, and Gneiting et al. (2007) suggest that this spread should be minimized, subject to achieving the optimal coverage.

To rank and compare the competing forecast distributions, it is useful to employ an objective measure of forecast performance. For this purpose, several scoring rules have been proposed. A scoring rule maps the predictive distribution and its corresponding observation to a numerical value, thereby objectively quantifying the forecast accuracy. Such a scoring rule is said to be proper if its statistical expectation is minimized when the forecast distribution is equivalent to the distribution from which the observations arose (Gneiting and Raftery 2007). Therefore, if a forecaster has access to the generation mechanism behind the observations, then there is no incentive for them to issue anything else as their forecast.

Both the logarithmic score and the CRPS are examples of proper scoring rules. The logarithmic score is a local score and thus relies on the predictive density function of the forecast, which is not readily available for forecasts in the form of an ensemble. Therefore, although Tödter and Ahrens (2012) propose a continuous ranked extension of the logarithmic score that is applicable to ensemble forecasts, the CRPS is more commonly implemented, since it reduces conveniently to

$$\mathrm{CRPS}(\mathbf{x}, y) = \frac{1}{M}\sum_{j=1}^{M}|y - x_j| - \frac{1}{2M^2}\sum_{j=1}^{M}\sum_{k=1}^{M}|x_j - x_k|, \qquad (16)$$

for an ensemble forecast $\mathbf{x}$ with M members. The CRPS is negatively oriented, so that a lower CRPS indicates a more skillful forecast. The CRPS thus rewards spread among the ensemble members while penalizing any deviation between the ensemble members and the observation, thereby assimilating both the reliability and sharpness of the forecast (Gneiting and Raftery 2007). The total CRPS is then taken to be the average CRPS over all forecasts.

The continuous ranked probability skill score (CRPSS) is also applied here to assess the difference in accuracy between the various prediction schemes. This skill score is calculated as the difference between the total CRPS for a reference forecast scheme and for a competing scheme, divided by the score for the reference (e.g., Wilks 2019). The skill score is positively oriented and bounded above by one, with values below zero indicating that the forecast under consideration performs worse than the reference to which it is being compared.

Since the aim is to obtain a calibrated forecast field over the region of interest, the forecasts are evaluated using UKV model analyses. For computational efficiency, rather than assessing the forecasts at every grid point on the UKV model domain, we consider forecasts at every eighth latitudinal coordinate and every sixth longitudinal coordinate on the domain. Results in the following section have thus been calculated on a grid of roughly 10 000 locations over the United Kingdom. Moreover, to better understand the qualitative behavior of the forecasts, the performance of the forecasts relative to weather observations at 116 station locations over the United Kingdom is also illustrated, having bilinearly interpolated the forecast field to these sites. The synoptic stations considered here are displayed in Fig. 1.

5. Results

a. Gridded forecast performance

Rank histograms for the raw ensemble forecasts and those generated using the various postprocessing methods are displayed in Fig. 4 at a lead time of 36 h. The MOGREPS-UK ensemble prediction system in this case exhibits a pronounced negative bias, failing to capture the higher temperature observations; this is particularly pertinent in summer. Postprocessing using nonhomogeneous Gaussian regression trained using minimum CRPS estimation addresses this bias, though the observed temperature is found to lie outside the range of ensemble members more often than would be expected if the ensemble were calibrated, indicating the forecast is underdispersed and hence overconfident. Conversely, when the NGR approach is trained by minimizing the logarithmic score, the resulting forecasts become overdispersed, reflecting the higher penalty that the logarithmic score assigns to overconfident forecasts.

A similar result is presented in Gebetsberger et al. (2018). The authors therefore propose employing a similar postprocessing framework featuring distributions with heavier tails, such as the logistic distribution. However, when the CRPS is chosen as the loss function, the resulting forecasts are also found here to lack dispersion. When trained by minimizing the logarithmic score, on the other hand, both the logistic and skew-logistic NR approaches appear reasonably well calibrated, whereas the NGR forecasts applied to Yeo–Johnson transformed temperatures are slightly overdispersed. Conversely, this transformation-based approach appears to rectify deficiencies in the alternative methods when trained using minimum CRPS estimation, though there

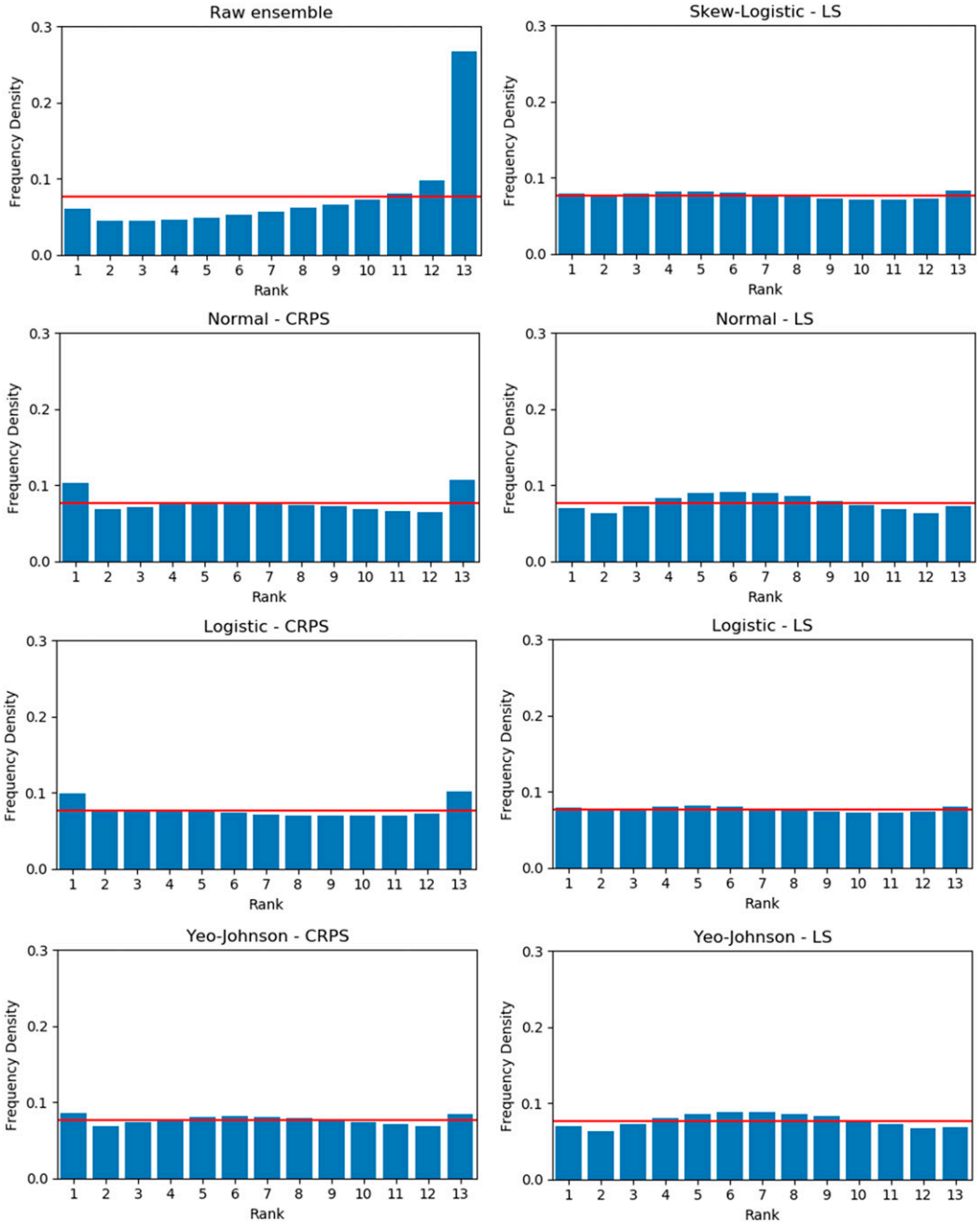

FIG. 4. Rank histograms for the raw ensemble forecasts and the various postprocessing methods, trained by (left) minimizing the CRPS and (right) minimizing the logarithmic score, LS (i.e., maximum likelihood) at a lead time of 36 h. The UKV model analyses are treated as the observed values and the ranks have been aggregated over all dates and locations. The horizontal red line shown at 1/13 is indicative of perfect calibration.

appears to be some remaining structure in the histogram that indicates a heavier-tailed forecast distribution may be more appropriate even after transformation. As such, given the large number of forecasts used to construct these rank histograms, a chi-squared test for uniformity (e.g., Wilks 2019) indicates that none of the postprocessing methods produce forecasts that are perfectly probabilistically calibrated.

The rank histograms in Fig. 4 are constructed using the UKV model analyses as the observed temperature values. Figure 5 shows the analogous histograms when verifying the ensembles against temperature recordings at synoptic stations over the United Kingdom. Since the postprocessing methods are trained using UKV model analyses, they are suited to address the discrepancies between the MOGREPS-UK

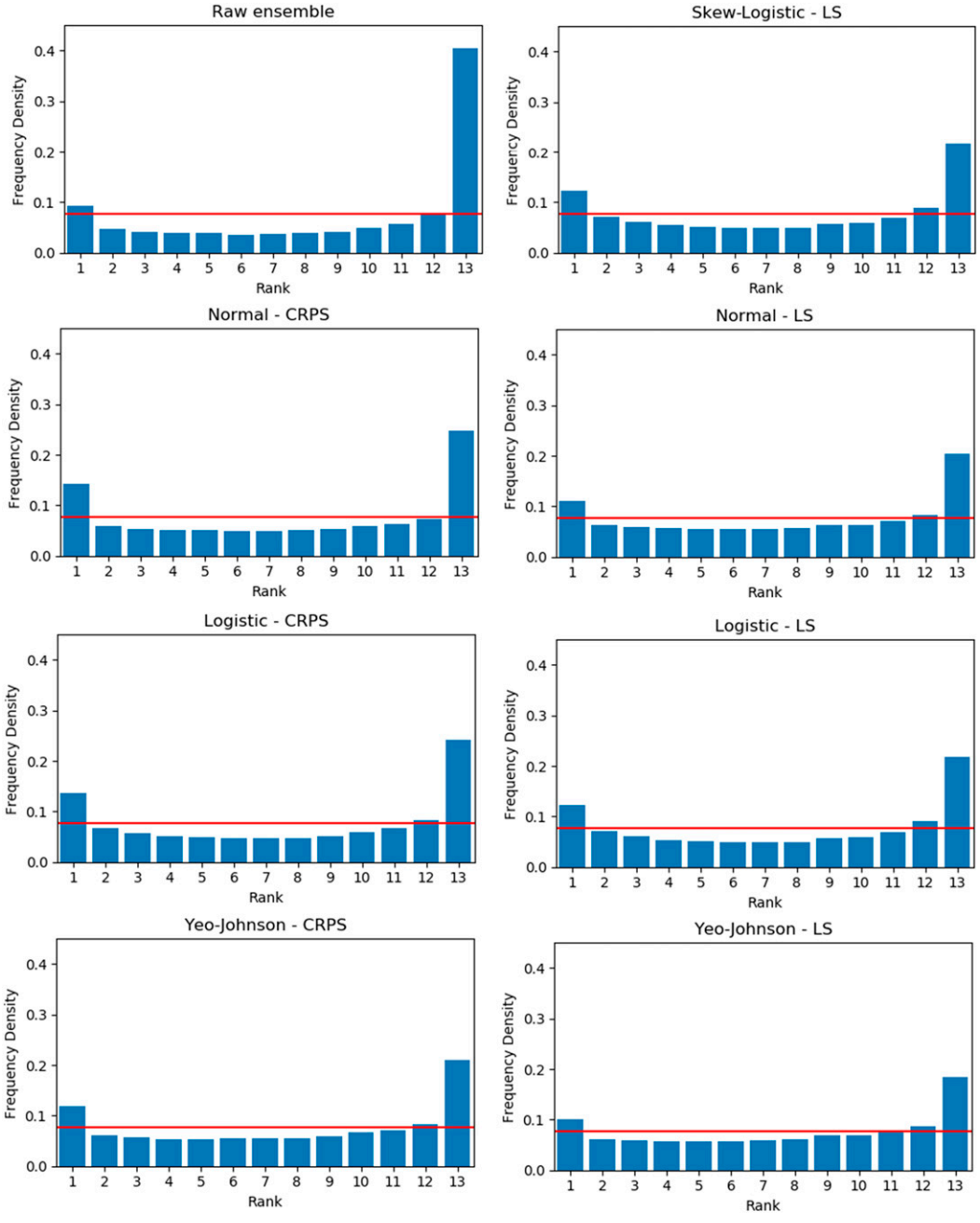

FIG. 5. As in Fig. 4, but with temperature recordings at synoptic stations treated as the observed values.

ensemble forecasts and these analysis fields. The postprocessing methods do not, however, represent the additional uncertainty that arises due to error in the analysis, and also that induced by downscaling the gridded forecasts to individual sites. As such, the rank histograms in Fig. 5 suggest the ensembles, even after postprocessing, are subject to considerable bias and dispersion errors, corroborating recent results in Feldmann et al. (2019). Nonetheless, the approaches based on Yeo–Johnson transformations reduce the underdispersion relative to the alternative methods.

The accuracy of the different postprocessing methods can be compared more formally using the CRPSS (again calculated across all locations and dates under consideration), available in Fig. 6. IMPROVER currently implements the NGR approach trained using minimum CRPS estimation, and this thus constitutes a canonical choice for the reference scheme when computing the skill scores. Results are shown for the forecasts evaluated against both the UKV model analyses and the station data. In the former case, the two postprocessing methods applied to Yeo–Johnson transformed

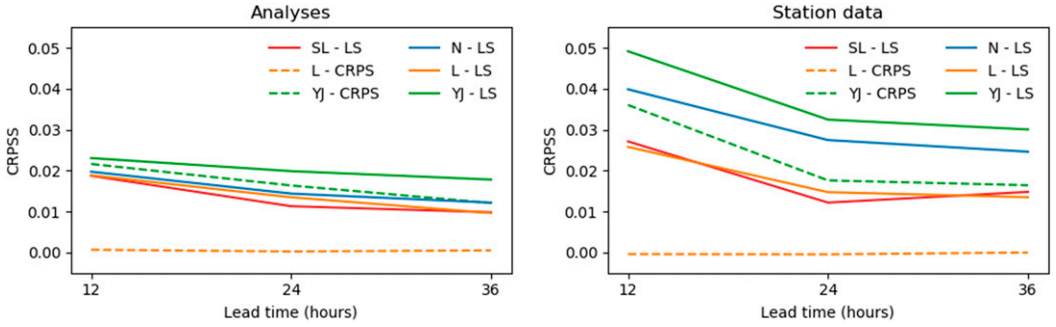

FIG. 6. The CRPSS for the different postprocessing methods at each lead time, relative to the NGR approach trained using minimum CRPS estimation. The forecasts are verified against (left) UKV model analyses and (right) weather station observations. The colors distinguish between the different parametric assumptions, while the line type reflects the loss function used to train the postprocessing methods. All standard errors in the first plot are negligible (see Table 2), while those in the second plot are all roughly 0.002 for all lead times. These have been omitted from the plot to aid interpretation.

forecasts provide the largest benefit, with skill scores roughly equal to 2% at all lead times, and the improvements are yet larger when the forecasts are assessed using station data, reaching almost 5% for forecasts 12 h in advance. Surprisingly, the approaches trained by minimizing the logarithmic score marginally outperform those designed to minimize the CRPS, even though forecast accuracy is assessed here using the CRPS.

We note, however, that this is not surprising when the forecasts are evaluated relative to station observations, since Fig. 4 illustrates that these approaches tend to be more overdispersed than those trained using minimum CRPS estimation. These methods thus inadvertently account for some of the additional uncertainty present when forecasting the less predictable station data, leading to an increased accuracy when assessed using observations of this type. More generally, this highlights that evaluating postprocessing methods against observations that exhibit markedly different characteristics to those used to train the methods can lead to invalid inferences regarding the quality of the postprocessing models. For example, even if a postprocessing method were capable of determining the exact generating process underlying the analysis fields in the training dataset, this method would not necessarily perform best when assessed using weather recordings at synoptic stations.

Of course, provided a suitable scoring rule is employed, the forecasts generated by the postprocessing methods can reliably be compared regardless of the choice of observations on which the comparison is based. But if the goal of the study is to gauge the effectiveness of various postprocessing methods, in terms of their ability to address the errors that they encounter in the training dataset, then the different methods should be evaluated using the same type of observations as those with which they are trained. On the other hand, of course, if the principal goal is to employ a postprocessing method that provides the most accurate forecasts at the synoptic weather stations, but practical constraints mean only model analyses are available to train the postprocessing methods, then it might be worthwhile to choose a postprocessing framework that is known to overpredict forecast uncertainty.

Since the interest here is on comparing the competing postprocessing methods, all results are henceforth presented when evaluating forecast performance against the UKV model analysis fields. Table 1, for example, displays the average CRPS over all locations at a lead time of 12 h. The mean squared error (MSE) of the ensemble mean forecast is also presented, as is the average range of the ensemble and the corresponding coverage. The CRPS for all postprocessing methods improves upon that of the raw ensemble by over 20%, while the Yeo–Johnson transformed predictive distributions generate further improvements upon the other approaches. The MSEs, on the other hand, are largely indistinguishable between the different postprocessing methods, suggesting the improvements are mainly due to a better representation of the shape of the predictive distribution. The coverage in this case is the proportion of instances in which the observed temperature falls within the ensemble members. Since the ensembles are each comprised of 12 members, the optimal coverage is 11/13 = 0.85. As was observed in the rank histograms, the normal and logistic distributions trained using minimum CRPS estimation are underdispersed, issuing coverages that fall below the optimal value. Estimating coefficients using maximum likelihood, or allowing the predictive distribution to exhibit skew, on the other hand, increases the spread of the ensemble members, and, in turn, produces forecasts that are well-calibrated with respect to this measure.

The extra flexibility in the skewed forecast distributions is attributable to the inclusion of an additional parameter: λ for the skew-logistic distribution and τ for the Yeo–Johnson transformation. A time series of this parameter, estimated for each forecast day in the test dataset, is shown in Fig. 7 for both day- and nighttime predictions. The distributions of these parameters indicate that at 1500 UTC, the conditional distribution of the temperature observations given the ensemble output is negatively skewed in winter ($\lambda < 1$, $\tau > 1$), and positively skewed in summer ($\lambda > 1$, $\tau < 1$). In autumn and spring, both coefficients are closer to one, suggesting the more parsimonious normal and logistic distributions are suffice during these seasons. For nighttime temperature

TABLE 1. CRPS, MSE, and the average width and coverage of 85% prediction intervals defined by the range of the ensemble members, with corresponding standard errors (scaled by 10^4) displayed in parentheses. Since the ensembles comprise 12 members, an optimal coverage would be 11/13 = 0.8462. All metrics have been computed at a lead time of 12 h using the UKV model temperature analyses as observations, and are averaged over all locations and days under consideration. The optimum CRPS, MSE, and coverage among the different methods is shown in boldface.

	CRPS	MSE	Width	Coverage
Raw ensemble	0.5096 (3)	0.7640 (8)	1.4372 (4)	0.6078 (2)
Normal–CRPS	0.3988 (2)	0.5408 (7)	1.7445 (4)	0.7780 (2)
Logistic–CRPS	0.3985 (2)	0.5408 (7)	1.7927 (5)	0.7875 (2)
Yeo–Johnson–CRPS	0.3902 (2)	0.5429 (7)	1.8821 (3)	0.8367 (2)
Normal–LS	0.3909 (2)	0.5415 (7)	2.0197 (2)	0.8654 (2)
Logistic–LS	0.3913 (2)	**0.5406 (7)**	1.8965 (3)	**0.8453 (2)**
Yeo–Johnson–LS	**0.3896 (2)**	0.5440 (7)	2.0851 (4)	0.8714 (2)
Skew-logistic–LS	0.3913 (2)	0.5410 (7)	1.8891 (2)	0.8439 (2)

forecasts, there is less variation in the shape coefficients, and the predictive distributions appear slightly negatively skewed throughout the year. These results are reinforced by the continuous ranked probability skill score (CRPSS), which is used here to measure the improvement of the various NR approaches relative to the normal forecast distribution trained by minimizing the CRPS. The CRPSS, displayed separately for each season in Table 2, indicates that improvements are largest in summer and winter, though still noticeable in autumn and spring. The large negative bias in the raw MOGREPS-UK ensemble forecasts is also apparent in Table 2, with statistical postprocessing offering the most benefit in spring and summer. Analogous conclusions are drawn when verifying the forecasts against the station data (not shown).

Perhaps surprisingly, although the forecasts benefit from the increased flexibility provided by the Yeo–Johnson transformation, the skew-logistic forecast distributions perform comparatively to the logistic NR approach, both quantitatively and qualitatively. This is in part explained by the upper bound on the positive skewness of these forecast distributions, as discussed in section 2, which can be seen from the fluctuating behavior of the shape parameter during summer in Fig. 7. This is further reinforced by the seasonal skill scores in Table 2, where the quantitative performance of the skew-logistic forecasts in summer is almost identical to that of the original logistic forecasts.

Last, not only does the skewness of the unconditional temperature distribution change for different seasons, but Fig. 2 indicates that it varies also for different locations. Figure 8 therefore displays the CRPSS for the forecasts generated using NGR applied to Yeo–Johnson transformed temperatures, relative to those using conventional NGR, calculated separately for each grid point. Both methods are trained here by minimizing the CRPS over the training data, allowing focus to be placed on the benefits gained by the more flexible distribution. The Yeo–Johnson-based postprocessing approach performs marginally worse than NGR at a band of locations over the North Atlantic and the North Sea, but significantly improves the performance of the resulting forecasts at inland locations across the United Kingdom. The accuracy of forecasts at individual grid points improves by as much as 10%, with the largest benefits appearing in mountainous regions in northern Scotland, which agrees with results in Gebetsberger et al. (2019). Schuhen et al. (2020) have also recently identified deficiencies in the MOGREPS-UK output and associated NGR postprocessed forecasts when predicting the temperature at mountainous locations.

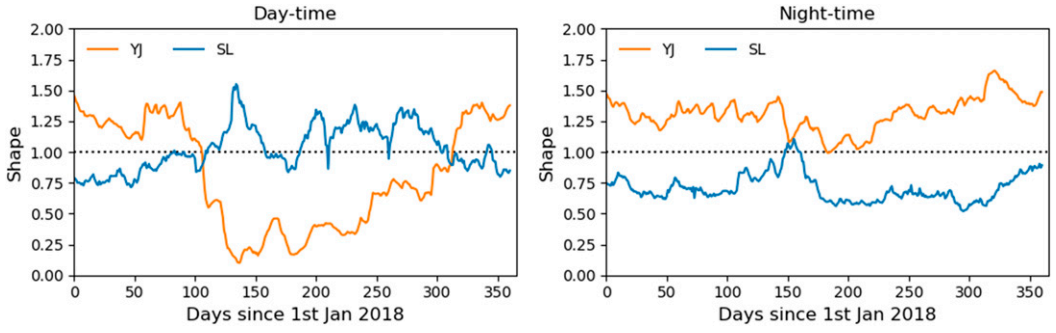

FIG. 7. Time series of the shape coefficient of the Yeo–Johnson transformation τ and skew-logistic predictive distributions λ as estimated over a time-adaptive 30-day training window. The shape is displayed for both (left) daytime (1500 UTC; 12-h forecasts) and (right) nighttime (0300 UTC; 24-h forecasts) temperatures. The shapes corresponding to 36-h forecasts are very similar to those displayed for 12-h forecasts.

TABLE 2. CRPSS (scaled by 100) for the prediction systems relative to nonhomogeneous Gaussian regression trained using minimum CRPS estimation, displayed separately for each season. Standard errors (computed using 1000 nonparametric bootstrap resamples and scaled by 100) are displayed in parentheses next to the score. The skill scores have been computed at a lead time of 12 h using the UKV model temperature analyses as observations, and are averaged over all locations and days under consideration. The optimum skill score in each season among the different methods is shown in boldface.

	Autumn	Spring	Summer	Winter	Total
Raw ensemble	−19.24 (0.07)	−32.73 (0.09)	−42.73 (0.13)	−10.57 (0.05)	−27.81 (0.04)
Logistic–CRPS	−0.10 (0.00)	0.12 (0.00)	0.04 (0.00)	−0.01 (0.00)	0.07 (0.00)
Yeo–Johnson–CRPS	1.12 (0.01)	1.01 (0.02)	**3.29 (0.02)**	3.18 (0.01)	2.16 (0.01)
Normal–LS	**1.47 (0.01)**	**1.65 (0.01)**	1.94 (0.01)	2.92 (0.01)	1.97 (0.01)
Logistic–LS	1.39 (0.01)	1.15 (0.01)	2.38 (0.01)	2.67 (0.01)	1.89 (0.01)
Yeo–Johnson–LS	1.39 (0.01)	1.22 (0.01)	3.14 (0.02)	**3.49 (0.02)**	**2.31 (0.01)**
Skew-logistic–LS	1.32 (0.01)	1.22 (0.01)	2.38 (0.01)	2.59 (0.01)	1.88 (0.01)

b. *Local forecast performance*

Based on Fig. 8, a suitable extension of the postprocessing framework implemented herein would be to calibrate grid points over land and sea separately. More generally, since the postprocessing methods are trained using temperature forecasts and observations aggregated across all (over 500 000) grid points on the UKV model domain, it could be the case that the error distribution of the ensemble mean forecast becomes skewed due to the combination of (potentially symmetric) forecast error distributions across several locations. In this respect, although the previous section demonstrated that skewed predictive distributions can help to account for this behavior, the benefits of these approaches may diminish if spatial information were incorporated into the postprocessing models.

To investigate whether or not this is the case, we restrict attention to the temperature forecasts at 116 grid points over the UKV domain, which correspond to the grid points associated with the station locations displayed in Fig. 1. The postprocessing set up is largely similar to before: all methods considered in the previous section are also compared here, trained using both maximum likelihood and minimum CRPS estimation over a rolling 30-day window, with the UKV temperature analysis fields still treated as the observations. However, in contrast to the previous setting, a local postprocessing framework is applied, whereby the coefficients of the various postprocessing methods are estimated separately at each of the 116 grid points. In addition, results are also presented here for the skew-logistic NR approach trained using minimum CRPS estimation, which can feasibly be implemented using the reduced amount of training data; the training dataset now consists of only 30 forecast–observation pairs. Details regarding how this approach is implemented are discussed in the appendix.

Figure 9 displays the rank histograms for the two postprocessing methods that employ a normal predictive distribution in this localized setting, at a lead time of 36 h. A similar pattern manifests to that observed in Fig. 4: the forecasts trained by minimizing the CRPS are noticeably underdispersed, whereas those trained using maximum likelihood overestimate the forecast uncertainty. This is the case also for the logistic NR approaches (not shown), suggesting even when grid points are considered individually, there is still the need to make more flexible parametric assumptions when postprocessing.

Boxplots of the CRPS for all postprocessing methods, averaged over all locations and test days, are presented in Fig. 10 for the same lead time. First, we note that when applying a local postprocessing approach, the discrepancy between the predictability of model analyses and station recordings is still apparent. The CRPS, which is defined on the same scale as the temperature values, is larger for forecasts assessed using the weather station data, reiterating that the postprocessing methods are less adept at capturing the station-specific temperatures than they are at predicting the UKV analyses. This is also true for the raw MOGREPS-UK output. In both cases, however, all postprocessing methods offer substantial improvements upon the raw, uncorrected ensemble forecast, as expected. The postprocessing methods in Fig. 10 have been ordered according to their median CRPS value, which is lowest for the two methods that employ a Yeo–Johnson transformation, regardless of whether UKV analyses or station recordings have been used to assess the forecasts.

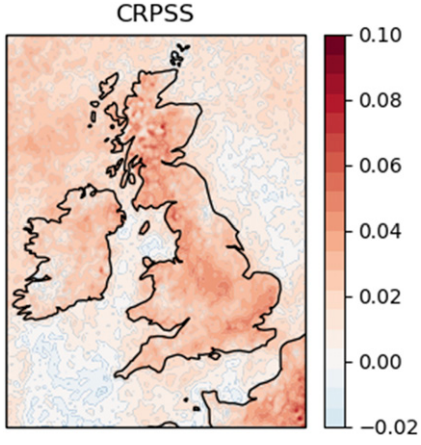

FIG. 8. Map of the continuous ranked probability skill score (CRPSS) for the Yeo–Johnson transformed NGR approach, relative to the standard NGR forecasts, estimated for each grid point over all of 2018 at a lead time of 12 h. Both methods have been trained using minimum CRPS estimation, and the gridded UKV model analysis fields are treated as the observations.

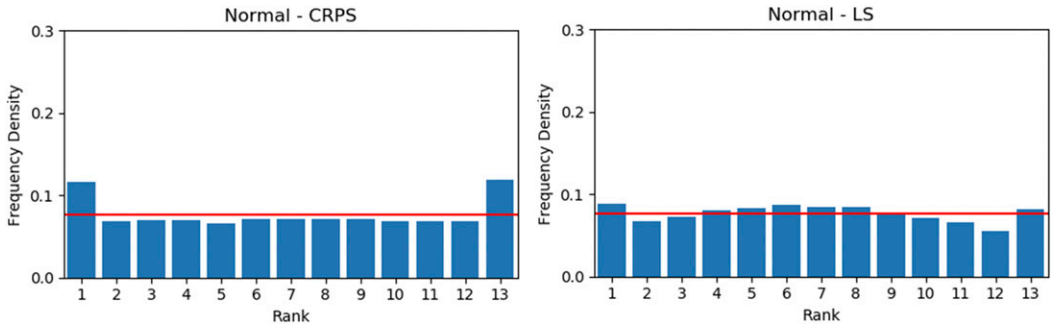

FIG. 9. Rank histogram for the local NGR forecasts trained using (left) minimum CRPS estimation and (right) maximum likelihood at a lead time of 36 h. The UKV model analyses are treated as the observed values and the ranks have been aggregated over all dates and locations. The horizontal red line shown at 1/13 is indicative of perfect calibration.

The corresponding skill scores for all methods are displayed in Table 3, with the NGR approach trained using minimum CRPS estimation again chosen as the reference scheme. Although Fig. 9 suggests that the assumption of normality is invalid, the skill scores indicate that the alternative NR approaches offer little improvement upon this baseline approach. The reason for this appears to be a result of the more flexible postprocessing methods becoming overfit on the reduced training dataset, since they require the estimation of an additional shape parameter. As such, there are a few forecast cases in which these approaches perform particularly poorly in comparison with the reference approach, resulting in heavier tailed distributions of the CRPS values (see Fig. 10). Hence, although the median CRPS value of the Yeo–Johnson-based approaches is lower than that of the reference scheme, the mean is higher, leading to negative skill scores. This sensitivity to the amount of training data is particularly pertinent for the skew-logistic approach, since the shape coefficient is estimated simultaneously to the other postprocessing parameters. The postprocessing approach applied to Yeo–Johnson transformed temperatures, on the other hand, could more easily be adapted to account for the amount of training data by estimating the shape coefficient over an augmented dataset, possibly utilizing information from several locations. The remaining postprocessing parameters could then be estimated locally, after obtaining a more reliable estimate for τ.

6. Discussion

This paper has studied the performance of short-range temperature forecast fields over the United Kingdom, issued by the Met Office's MOGREPS-UK ensemble prediction system. The MOGREPS-UK forecasts exhibit a strong negative bias, and statistical postprocessing is therefore necessary to recalibrate the numerical model output. To do so, a nonhomogeneous regression approach is implemented here with four different choices of the parametric assumptions. Focus is particularly on the performance of skewed predictive distributions, including a variant of the logistic distribution that has recently been proposed to account for changes in the shape of empirical temperature distributions (Gebetsberger et al. 2019), as well as a novel approach that nonlinearly transforms the temperature forecasts prior to postprocessing, which can similarly generate asymmetric predictive distributions.

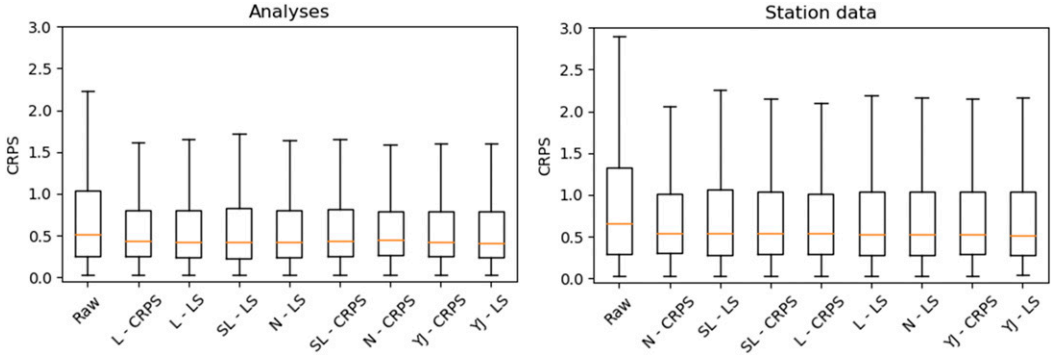

FIG. 10. Boxplots of the CRPS for the raw ensemble forecast and the various local postprocessing methods (as described in section 5b) verified against (left) UKV model analyses and (right) weather station observations at a lead time of 36 h. The boxes contain the median (orange line) and the lower and upper quartiles of the empirical CRPS distribution. Values of the CRPS that exceed (fall below) the upper (lower) quartile plus (minus) 1.5 times the interquartile range are defined as outliers, and have been removed from both plots. The methods have been ordered by decreasing median CRPS.

TABLE 3. CRPSS (scaled by 100) for the prediction systems relative to nonhomogeneous Gaussian regression trained using minimum CRPS estimation, when a localized postprocessing method is implemented. Standard errors (computed using 1000 nonparametric bootstrap resamples and scaled by 100) are displayed in parentheses next to the score. The skill score is shown at all lead times using the UKV model temperature analyses as observations, and is averaged over all grid points and days under consideration. The optimum skill score among the different methods is shown in boldface.

	12 h	24 h	36 h
Raw ensemble	−42.01 (0.61)	−12.22 (0.31)	−21.72 (0.42)
Logistic–CRPS	0.30 (0.18)	0.30 (0.11)	0.01 (0.09)
Yeo–Johnson–CRPS	−0.04 (0.08)	−0.63 (0.15)	−0.80 (0.18)
Skew-logistic–CRPS	−0.17 (0.18)	−2.31 (0.15)	−2.31 (0.14)
Normal–LS	−0.83 (0.13)	−0.79 (0.14)	−0.49 (0.11)
Logistic–LS	**0.32 (0.20)**	**0.47 (0.12)**	**0.10 (0.06)**
Yeo–Johnson–LS	−1.50 (0.14)	−1.00 (0.17)	−0.87 (0.18)
Skew-logistic–LS	−1.65 (0.23)	−2.06 (0.18)	−1.78 (0.18)

Regardless of the choice of parametric family that is employed in nonhomogeneous regression, postprocessing yields forecasts that are significantly more accurate and reliable than the raw temperature ensemble forecasts, particularly in summer.

However, it is common to employ a normal distribution within the NR framework when postprocessing temperature forecasts, whereas such an approach is found here to be inappropriate. In particular, the resulting forecasts are found to be either under or overdispersed, depending on the loss function used to train the forecasts, corroborating results in Gebetsberger et al. (2018). Instead, the most accurate forecasts, as measured using the continuous ranked probability score, are generated by an approach that applies nonhomogeneous Gaussian regression after having transformed the temperature forecasts and observations in the training dataset so that they appear more symmetric. This is true when using both high-resolution UKV model analyses and station data to assess the resulting forecasts, though we argue that conclusions should be treated with caution in the latter case, since these observations are less predictable than the analysis fields, meaning postprocessing methods that overestimate the predictive uncertainty appear more appealing. The nonlinear transformation implemented here is the Yeo–Johnson transformation (Yeo and Johnson 2000), which belongs to the more general class of power transformations frequently used in the wider field of statistical modeling (Wilks 2019). As such, although applied here to temperature forecasts, power transformations could also easily be implemented when postprocessing several alternative weather variables (Hemri et al. 2015).

In any case, as in Gebetsberger et al. (2019), the results presented herein demonstrate the potential benefit provided by more flexible parametric assumptions when postprocessing temperature forecasts. We illustrate this when applying a global postprocessing approach, whereby all grid points are recalibrated simultaneously, but demonstrate that deficiencies in conventional methods exist also in more local settings. However, as when incorporating additional predictors into the postprocessing model, more complex predictive distributions render the postprocessing methods more dependent on the amount of available training data. Moreover, it may be the case that the more flexible parametric assumptions add information to the forecast that could alternatively be introduced via additional predictors, and future studies may wish to investigate this. Nonetheless, as long as the unconditional temperature distribution exhibits skew, we anticipate that asymmetric predictive distributions will be beneficial at longer lead times than those considered here; as the lead time increases, the observed weather variable becomes independent of the inputs to the postprocessing model, meaning the conditional distribution of the weather variable reverts to its climatological, or unconditional distribution (Allen et al. 2020).

Furthermore, all forecasts in this study have been evaluated using both UKV model analyses, as well as temperature recordings at synoptic stations over the United Kingdom. Assessing the forecasts against model analyses allows the spatial characteristics of forecast performance to be better understood; as in Gebetsberger et al. (2019), the benefits of issuing skewed predictive distributions were particularly large in mountainous regions, with improvements at individual grid points reaching almost 10%. Similar improvements were also observed when verifying forecasts against temperature recordings at synoptic stations over the United Kingdom. Although there is still error in these recordings (Ferro 2017), they generally provide a much more accurate reflection of the weather that actually materializes. However, since the postprocessing methods are trained against high-resolution model analyses, they are designed to correct forecast biases relative to these gridded analysis fields. As such, the postprocessing methods are poorly suited to capture the additional uncertainty present when predicting the station-based temperature recordings, resulting in underdispersed forecast distributions. Rank histograms suggest that this underdispersion is less acute for the approaches that overpredict the uncertainty in the training data, and this is reflected by measures of forecast accuracy. The results presented herein therefore call for more effective approaches of combining the station data and the analysis fields when postprocessing. This could be achieved, for example, by treating the postprocessed predictive distributions trained using the analysis fields as prior distributions when forecasting the station data, or by suitably assimilating the two sources of information prior to fitting the postprocessing model.

Finally, this study has considered forecast distributions constructed using the nonhomogeneous regression (NR) framework, which generally relies on specifying a unimodal

predictive distribution centered around the (bias-corrected) ensemble mean, and with scale or variance that depends on the ensemble spread. One benefit of the skew-logistic forecast distribution is that the shape parameter of these forecast distributions could similarly be estimated using the ensemble sample skewness as a predictor, to incorporate flow-dependent shape information provided by the numerical model output. This was not considered here to maintain comparison with alternative approaches, though this could easily be investigated in future studies. Alternatively, it might be the case that if the ensemble members naturally reflect the skew in the temperature distributions, then a postprocessing approach that dresses each ensemble member individually, such as Bayesian model averaging (BMA; Raftery et al. 2005), might be able to utilize symmetric component distributions while also capturing this asymmetry. This reflects the additional flexibility provided by the mixture distribution used in BMA compared to NR, since it uses information independently regarding each ensemble member. In this sense, postprocessing methods that make simple assumptions about the conditional distribution of the weather variable being forecast are at times inadequate. Suitably transforming the data or utilizing more flexible parametric distributions (e.g., Allen et al. 2019) are potential ways of alleviating this, as might be non- or semiparametric approaches, which have recently received increased attention in the field of postprocessing (Van Schaeybroeck and Vannitsem 2015; Taillardat et al. 2016; Henzi et al. 2020; Bremnes 2020).

Acknowledgments. Sam Allen was supported during this work by a NERC Industrial CASE studentship under grant reference NE/N008693/1. The authors thank Chris Ferro, Peter Challenor, Stéphane Vannitsem, and Richard Woesler for their valuable input, and three anonymous reviewers, whose suggestions have significantly improved the quality of this manuscript.

Data availability statement. The postprocessing of MOGREPS-UK temperature forecasts was performed within IMPROVER, a postprocessing suite currently under development at the Met Office, available at https://github.com/metoppv/improver. The MOGREPS-UK temperature forecasts and UKV analysis fields were extracted from the Met Office's Managed Archive Storage System (MASS).

APPENDIX

Minimum CRPS Estimation with the Type-I Generalized Logistic Distribution

In this appendix we derive expressions of the continuous ranked probability score (CRPS) for forecasts in the form of Type-I generalized logistic distributions, with probability density function (PDF) and cumulative distribution function (CDF) as defined in Eqs. (3) and (4). Let F_λ denote the CDF of the standard skew-logistic distribution $L(0, 1, \lambda)$. The CRPS for forecasts in this form is defined as

$$\begin{aligned}\mathrm{CRPS}[L(0,1,\lambda),y] &= \int_{-\infty}^{\infty} [F_\lambda(u) - \mathbb{1}(u \geq y)]^2\, du \\ &= \int_{-\infty}^{y} F_\lambda^2(u)\, du + \int_{y}^{\infty} [1 - F_\lambda(u)]^2\, du. \end{aligned} \tag{A1}$$

Without loss of generality, we restrict attention to the standard skew-logistic distribution, with $\mu = 0$ and $\sigma = 1$, since the CRPS in this case can easily be extended for other location and scale parameters using Eq. (15). Note now that the CDF of the standard generalized logistic distribution, F_λ, is simply the standard logistic CDF, F_L, raised to the power of the shape parameter λ. Substituting $s = F_L(u)$ gives $s^\lambda = F_\lambda(u)$ and $ds = s(1 - s)du$, so that Eq. (A1) becomes

$$\mathrm{CRPS}[L(0,1,\lambda),y] = \int_0^{F_L(y)} \frac{s^{2\lambda-1}}{1-s}\, ds + \int_{F_L(y)}^{1} \frac{(1-s^\lambda)^2}{s(1-s)}\, ds. \tag{A2}$$

If the shape parameter λ is a rational number, that is, $\lambda = a/b$ with $a, b \in \mathbb{N}$, then the CRPS is available in closed-form. Let $v = F_L^{1/b}(u)$ so that $v^a = F_\lambda(u)$ and $dv = [v\,(1 - v^b)/b]du$. Then Eq. (A2) can be written as

$$\mathrm{CRPS}[L(0,1,a/b),y] = b\int_0^{F_L^{1/b}(y)} \frac{v^{2a-1}}{1-v^b}\, dv + b\int_{F_L^{1/b}(y)}^{1} \frac{(1-v^a)^2}{v(1-v^b)}\, dv. \tag{A3}$$

The two integrands are now rational functions and the integrals can be calculated analytically using partial fractions. For $b = 1$, we recover Eq. (A2) with $\lambda = a$, and for all $a \in \mathbb{N}$ we get

$$\begin{aligned}\mathrm{CRPS}[L(0,1,a),y] &= y - 2\log F_L(y) \\ &\quad + \sum_{k=1}^{a-1} \frac{1}{k}[1 - 2F_L^k(y)] - \sum_{k=a}^{2a-1} \frac{1}{k}. \end{aligned} \tag{A4}$$

For $a = 1$, this together with Eq. (15) gives Eq. (12) and, e.g., for $a = 2$ we get

$$\mathrm{CRPS}[L(0,1,2),y] = y + \frac{1}{6} - 2F_L(y) - 2\log F_L(y). \tag{A5}$$

For $b = 2$, the expression valid for all odd $a \in \mathbb{N}$ is still rather simple:

$$\begin{aligned}\mathrm{CRPS}[L(0,1,a/2),y] &= y + 4\log\frac{1 + F_L^{-1/2}(y)}{2} \\ &\quad + 4\sum_{k=0}^{(a-3)/2} \frac{1}{2k+1}[1 - F_L^{(2k+1)/2}(y)] \\ &\quad - \sum_{k=1}^{a-1} \frac{1}{k}. \end{aligned} \tag{A6}$$

For $a = 3$, for example, we get

$$\mathrm{CRPS}[L(0,1,3/2),y] = y + \frac{5}{2} - 4F_L^{1/2}(y) + 4\log\frac{1 + F_L^{-1/2}(y)}{2}. \tag{A7}$$

Furthermore, there is an infinite series representation of the CRPS for all positive real values of λ. Going back to Eq. (A2) and using the power series $1/(1-s) = \sum_{k=0}^{\infty} s^k$ we find

$$\begin{aligned}\mathrm{CRPS}[L(0,1,\lambda),y] = &\sum_{k=0}^{\infty}\int_0^{F_L(y)} s^{k+2\lambda-1}\,ds \\ &+\sum_{k=0}^{\infty}\int_{F_L(y)}^{1} s^{k-1}(1-2s^{\lambda}+s^{2\lambda})\,ds.\end{aligned} \tag{A8}$$

Integrating the terms of these series is straightforward, and the resulting components combine to produce

$$\begin{aligned}\mathrm{CRPS}[L(0,1,\lambda),y] = &-\log F_L(y) + \sum_{k=1}^{\infty}\frac{1}{k}[1-F_L^k(y)] \\ &-2\sum_{k=0}^{\infty}\frac{1}{k+\lambda}[1-F_L^{k+\lambda}(y)] \\ &+\sum_{k=0}^{\infty}\frac{1}{k+2\lambda}.\end{aligned} \tag{A9}$$

We remark that both the integration technique for rational λ and the infinite series technique are immediately applicable also to the CRPS in a couple of other settings. These are (i) the CRPS of a truncated skew-logistic distribution with any truncation point, (ii) the threshold-weighted CRPS (twCRPS; Gneiting and Ranjan 2011) of the skew-logistic distribution with any threshold (when using an indicator weight function) and (iii) the twCRPS of any truncated skew-logistic distribution. This might be useful in future work when postprocessing nonnegative meteorological variables such as wind speed or precipitation and/or evaluating the tail of a forecast distribution. Compact expressions for the CRPS and the twCRPS of the truncated logistic distribution ($\lambda = 1$) have been used by Allen et al. (2021) in the postprocessing of ensemble wind speed forecasts.

However, it is not immediately obvious how to efficiently utilize these expressions when numerically optimizing the CRPS for a skew-logistic distribution. One approach would be to employ symbolic algebra packages to evaluate the CRPS of the skew-logistic distribution analytically for a sequence of rational shape parameters, at a range of possible values of y. Interpolating this output would then provide a smooth function that approximates the CRPS at values of λ and y. Then, using Eq. (15), numerical optimization routines could be used to optimize the smooth interpolant with respect to the location, scale, and shape parameters over the training dataset.

Alternatively, numerical optimization routines could use finite approximations of the infinite series in Eq. (A9). However, the repeated evaluation of this series is more time consuming than computing the CRPS for normal and logistic distributions in Eqs. (11) and (12). This is especially true when large volumes of data are considered, as is the case here, since the convergence of the series is slow when observations in the training data lie in the extreme upper tail of the forecast distribution, which is more likely to occur when considering larger archives of data. To illustrate this, Fig. A1 displays the CRPS for a standard skew-logistic distribution with shape parameter equal to one-half, approximated using the infinite series in Eq. (A9) truncated at term K. Even with $K = 500$, the series approximation is not accurate in the extreme upper tail, and a considerably larger number of terms is required to avoid this.

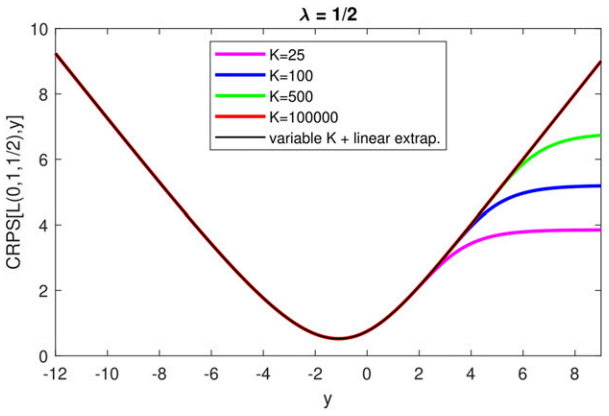

FIG. A1. CRPS of a standard skew-logistic predictive distribution with shape parameter equal to 1/2, plotted as a function of the observation y. The CRPS is approximated using the infinite-series representation in Eq. (13) truncated at K terms, for various choices of K.

Increasing the number of terms in the series obviously increases the time it takes to approximate the CRPS, prohibiting its use in numerical optimization routines. To circumvent this, we introduce an approximation of the infinite series in Eq. (A9) that uses a variable number of terms K depending on the value of y:

$$K = \begin{cases} 25, & y \le 1.5, \\ 100, & 1.5 < y \le 3, \\ 500, & 3 < y \le 5. \end{cases} \tag{A10}$$

If $y > 5$, then we make use of the linearity of the CRPS for large y, and approximate $\mathrm{CRPS}[L(0,1,\lambda),y]$ using $\mathrm{CRPS}[L(0,1,\lambda),5] + y - 5$, where $\mathrm{CRPS}[L(0,1,\lambda),5]$ is evaluated using the series with 500 terms. Figure A1 illustrates that even though at most 500 terms are used in the series using with approach, the resulting approximation of the CRPS performs just as well as that obtained using 100 000 terms in the series without employing a linear extrapolation in the upper tail. Hence, the optimization of the skew-logistic forecast distributions in section 5b has been performed using this approximation to the CRPS.

REFERENCES

Allen, S., C. A. T. Ferro, and F. Kwasniok, 2019: Regime-dependent statistical post-processing of ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **145**, 3535–3552, https://doi.org/10.1002/qj.3638.

——, ——, and ——, 2020: Recalibrating wind speed forecasts using regime-dependent ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **146**, 2576–2596, https://doi.org/10.1002/qj.3806.

——, G. R. Evans, P. Buchanan, and F. Kwasniok, 2021: Incorporating the North Atlantic Oscillation into the post-processing of

MOGREPS-G wind speed forecasts. *Quart. J. Roy. Meteor. Soc.*, **147**, 1403–1418, https://doi.org/10.1002/qj.3983.

Alley, R. B., K. A. Emanuel, and F. Zhang, 2019: Advances in weather prediction. *Science*, **363**, 342–344, https://doi.org/10.1126/science.aav7274.

Azzalini, A., 1985: A class of distributions which includes the normal ones. *Scand. J. Stat.*, **12**, 171–178.

Barnes, C., C. M. Brierley, and R. E. Chandler, 2019: New approaches to postprocessing of multi-model ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **145**, 3479–3498, https://doi.org/10.1002/qj.3632.

Box, G. E., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26**, 211–243.

Bremnes, J. B., 2020: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, **148**, 403–414, https://doi.org/10.1175/MWR-D-19-0227.1.

Dabernig, M., I. Schicker, A. Kann, Y. Wang, and M. N. Lang, 2020: Statistical post-processing with standardized anomalies based on a 1 km gridded analysis. *Meteor. Z.*, **29**, 265–275, https://doi.org/10.1127/metz/2020/1022.

Evans, G. R., and Coauthors, 2020: metoppv/improver: IMPROVER: A library of algorithms for meteorological post-processing (version 0.10.0).GitHub, https://doi.org/10.5281/zenodo.3744431.

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, https://doi.org/10.1175/MWR-D-14-00210.1.

——, D. S. Richardson, and T. Gneiting, 2019: Grid-versus station-based postprocessing of ensemble temperature forecasts. *Geophys. Res. Lett.*, **46**, 7744–7751, https://doi.org/10.1029/2019GL083189.

Ferro, C. A. T., 2017: Measuring forecast performance in the presence of observation error. *Quart. J. Roy. Meteor. Soc.*, **143**, 2665–2676, https://doi.org/10.1002/qj.3115.

Friedli, L., D. Ginsbourger, and J. Bhend, 2021: Area-covering postprocessing of ensemble precipitation forecasts using topographical and seasonal conditions. *Stochastic Environ. Res. Risk Assess.*, **35**, 215–230, https://doi.org/10.1007/s00477-020-01928-4.

Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, 2018: Estimation methods for nonhomogeneous regression models: Minimum continuous ranked probability score versus maximum likelihood. *Mon. Wea. Rev.*, **146**, 4323–4338, https://doi.org/10.1175/MWR-D-17-0364.1.

——, R. Stauffer, G. J. Mayr, and A. Zeileis, 2019: Skewed logistic distribution for statistical temperature post-processing in mountainous areas. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **5**, 87–100, https://doi.org/10.5194/ascmo-5-87-2019.

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., 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.

——, and R. Ranjan, 2011: Comparing density forecasts using threshold-and quantile-weighted scoring rules. *J. Bus. Econ. Stat.*, **29**, 411–422, https://doi.org/10.1198/jbes.2010.08110.

——, A. E. Raftery, A. H. Westveld III, 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.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gupta, R. D., and D. Kundu, 2010: Generalized logistic distributions. *J. Appl. Stat. Sci.*, **18**, 51–66.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, 2018: Practical aspects of statistical postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 187–217.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. national blend of models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Hemri, S., D. Lisniak, and B. Klein, 2015: Multivariate postprocessing techniques for probabilistic hydrological forecasting. *Water Resour. Res.*, **51**, 7436–7451, https://doi.org/10.1002/2014WR016473.

Henzi, A., G.-R. Kleger, and J. F. Ziegel, 2020: Distributional (single) index models. arXiv preprint arXiv:2006.09219.

Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions*. John Wiley & Sons, Ltd., 714 pp.

Jordan, A., F. Krüger, and S. Lerch, 2017: Evaluating probabilistic forecasts with scoring rules. *J. Stat. Software*, **90**, 1–37, https://doi.org/10.18637/jss.v090.i12.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Klein, N., and Coauthors, 2015: Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Stat.*, **9**, 1024–1052, https://doi.org/10.1214/15-AOAS823.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682, https://doi.org/10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2.

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

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1.

——, ——, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1.

Möller, A., and J. Groß, 2020: Probabilistic temperature forecasting with a heteroscedastic autoregressive ensemble postprocessing model. *Quart. J. Roy. Meteor. Soc.*, **146**, 211–224, https://doi.org/10.1002/qj.3667.

Oliver, H. J., M. Shin, and O. Sanders, 2018: Cylc: A workflow engine for cycling systems. *J. Open Source Software*, **3**, 737, https://doi.org/10.21105/joss.00737.

Oliver, H., and Coauthors, 2019: Workflow automation for cycling systems: The Cylc workflow engine. *Comput. Sci. Eng.*, **21**, 7–21, https://doi.org/10.1109/MCSE.2019.2906593.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast

ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

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 L. Büermann, 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. *J. Roy. Stat. Soc.*, **63**, 405–422, https://doi.org/10.1111/rssc.12040.

——, and G. König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 2582–2590, https://doi.org/10.1002/qj.2323.

——, 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 D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9**, 1328–1349, https://doi.org/10.1214/15-AOAS843.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.*, **140**, 3204–3219, https://doi.org/10.1175/MWR-D-12-00028.1.

——, T. Thorarinsdottir, and A. Lenkoski, 2020: Rapid adjustment and post-processing of temperature forecast trajectories. *Quart. J. Roy. Meteor. Soc.*, **146**, 963–978, https://doi.org/10.1002/qj.3718.

Siegert, S., P. G. Sansom, and R. M. Williams, 2016a: Parameter uncertainty in forecast recalibration. *Quart. J. Roy. Meteor. Soc.*, **142**, 1213–1221, https://doi.org/10.1002/qj.2716.

——, D. B. Stephenson, P. G. Sansom, A. A. Scaife, R. Eade, and A. Arribas, 2016b: A Bayesian framework for verification and recalibration of ensemble forecasts: How uncertain is NAO predictability? *J. Climate*, **29**, 995–1012, https://doi.org/10.1175/JCLI-D-15-0196.1.

Sloughter, J. M., 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.

Stephenson, D., C. Coelho, F. Doblas-Reyes, and M. Balmaseda, 2005: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions. *Tellus*, **57A**, 253–264, https://doi.org/10.3402/tellusa.v57i3.14664.

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.

Tang, Y., H. W. Lean, and J. Bornemann, 2013: The benefits of the Met Office variable resolution NWP model for forecasting convection. *Meteor. Appl.*, **20**, 417–426, https://doi.org/10.1002/met.1300.

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.*, **173**, 371–388, https://doi.org/10.1111/j.1467-985X.2009.00616.x.

——, and N. Schuhen, 2018: Verification: Assessment of calibration and accuracy. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 155–186.

Tödter, J., and B. Ahrens, 2012: Generalization of the ignorance score: Continuous ranked version and its decomposition. *Mon. Wea. Rev.*, **140**, 2005–2017, https://doi.org/10.1175/MWR-D-11-00266.1.

Van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, https://doi.org/10.1002/qj.2397.

Von Storch, H., and F. W. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences*. 4th ed. Elsevier, 840 pp..

Williams, R., C. A. T. Ferro, and F. Kwasniok, 2014: A comparison of ensemble post-processing methods for extreme events. *Quart. J. Roy. Meteor. Soc.*, **140**, 1112–1120, https://doi.org/10.1002/qj.2198.

Yeo, I.-K., and R. A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954–959, https://doi.org/10.1093/biomet/87.4.954.