Received: 19 October 2021 | Revised: 16 March 2022 | Accepted: 21 April 2022 | Published on: 4 July 2022

DOI: 10.1002/qj.4305

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble

George C. Craig[1] | Matjaž Puh[1] | Christian Keil[1] | Kirsten Tempest[1] | Tobias Necker[2] | Juan Ruiz[3] | Martin Weissmann[2] | Takemasa Miyoshi[4]

[1]Meteorological Institute Munich, Ludwig-Maximilians-Universität, Munich, Germany

[2]Institut für Meteorologie und Geophysik, Universität Wien, Vienna, Austria

[3]Centro de Investigaciones del Mar y la Atmosfera, CIMA/CONICET-UBA, Buenos Aires, Argentina

[4]RIKEN Center for Computational Science, Kobe, Japan

Correspondence
George C. Craig, Meteorological Institute Munich,
Ludwig-Maximilians-Universität, 80333 Munich, Germany.
Email: george.craig@lmu.de

Funding Information
This work was carried out in Project A6 of the Collaborative Research Centre *Waves to Weather*, funded by the DFG (Deutsche Forschungsgemeinschaft) grant SFB/TR165, with additional support from the Hans-Ertel Centre for Weather Research, Data Assimilation Branch, funded by the BMVI (Federal Ministry of Transport and Digital Infrastructure, grant DWD2014P8).

Abstract

The errors in numerical weather forecasts resulting from limited ensemble size are explored using 1,000-member forecasts of convective weather over Germany at 3-km resolution. A large number of forecast variables at different lead times were examined, and their distributions could be classified into three categories: quasi-normal (e.g., tropospheric temperature), highly skewed (e.g. precipitation), and mixtures (e.g., humidity). Dependence on ensemble size was examined in comparison to the asymptotic convergence law that the sampling error decreases proportional to $N^{-1/2}$ for large enough ensemble size N, independent of the underlying distribution shape. The asymptotic convergence behavior was observed for the ensemble mean of all forecast variables, even for ensemble sizes less than 10. For the ensemble standard deviation, sizes of up to 100 were required for the convergence law to apply. In contrast, there was no clear sign of convergence for the 95th percentile even with 1,000 members. Methods such as neighborhood statistics or prediction of area-averaged quantities were found to improve accuracy, but only for variables with random small-scale variability, such as convective precipitation.

KEYWORDS

ensemble, forecast uncertainty, probability distribution

1 | INTRODUCTION

Weather is by nature unpredictable, and especially so on the convective scale where errors grow rapidly (Lorenz, 1969; Hohenegger *et al.*, 2006; Leoncini *et al.*, 2010; Clark *et al.*, 2010; Keil *et al.*, 2014; Craig *et al.*, 2021). But even when deterministic forecasts are inaccurate, useful information can often be obtained in the form of probabilities. In a highly non-linear system like the atmosphere, the probability distribution is often

complex in form and evolving in time, and as a result most meteorological services base their probabilistic predictions on ensembles of numerical forecasts.

An ensemble prediction system must represent several sources of forecast uncertainty. One important source is the uncertainty in the initial conditions, as represented by the data assimilation (DA) system (Stensrud *et al.*, 2009; Sun *et al.*, 2014). Current operational DA systems, whether variational or ensemble-based, are derived under the assumption that error distributions are Gaussian, and do not perform well when this assumption is not met (Evensen and van Leeuwen, 2000). This happens more frequently at smaller scales because of the rapid non-linear error evolution in convective weather (Zhang, 2005). Another major source of uncertainty is model error. There is little consensus regarding how this should be represented in the ensemble system, but convective-scale ensembles may include variations in model parameters to represent uncertainty due to systematic errors (e.g. Gebhardt *et al.*, 2011; Kühnlein *et al.*, 2014), or stochastic parameterizations to represent unresolved processes (e.g. Bouttier *et al.*, 2012; Jankov *et al.*, 2017; Rasp *et al.*, 2018; Hirt *et al.*, 2019; Sakradzija *et al.*, 2020).

In practice, one of the most important factors for the quality of the probabilistic forecast is the size of the ensemble, that is, the number of member forecasts used to construct the predicted distribution of a forecast variable. A large ensemble is required to accurately capture the shape of the distributions, especially in the case of rare outlier events and non-Gaussian behaviors such as multimodality or heavy tails (Bannister *et al.*, 2017). Because of computational power limitations, operational ensembles in numerical weather prediction (NWP) centres rarely have more than 50 members for global models (e.g. see table 1 in Leutbecher, 2019) and even fewer for limited area models (e.g. Gebhardt *et al.*, 2011; Bouttier *et al.*, 2012; Hagelin *et al.*, 2017; Schwartz *et al.*, 2017; Frogner *et al.*, 2019; Keil *et al.*, 2020). This is not enough to accurately represent the non-linear evolution of the forecast distributions (Leutbecher, 2019). Indeed, using an intermediate atmospheric general circulation model, Kondo and Miyoshi (2019) performed experiments with up to 10,240 ensemble members, and concluded that approximately 1,000 ensemble members were necessary to represent important features of the distributions, such as multimodality and probability of extreme events.

Less is known about the problem of under sampling in convective-scale NWP. Harnisch and Keil (2015) found that an increase of ensemble size from 20 to 40 members led to a more accurate analysis and more accurate 3-hr forecasts. Hagelin *et al.* (2017) showed a large improvement in forecast skill for precipitation with a doubling of ensemble size from 12 to 24 using the Met Office convective-scale ensemble (MOGREPS-UK). Raynaud and Bouttier (2017) compared the benefits of increasing ensemble size from 12 to 34 members to the benefits of increasing horizontal resolution from 2.5 to 1.3 km. They showed that the increase in ensemble size is more beneficial than a resolution increase for lead times greater than about 1 hr, when there is a larger uncertainty to be sampled. This result is consistent with Legrand *et al.* (2016), who found that non-Gaussianity, and hence the need for large samples, increases with forecast lead time. This, and previous studies using big ensembles at both global and regional scales, consistently show that the largest non-Gaussianity arises from highly non-linear processes in deep convective clouds (Miyoshi *et al.*, 2014; Jacques and Zawadzki, 2015). Recent studies using data assimilation in 1,000-member convective-scale ensembles have shown that non-Gaussianity can develop in much less than an hour in deep moist convection, originating in the region of the convective updrafts (Kawabata and Ueno, 2020; Ruiz *et al.*, 2021).

The preceding results all suggest that the ensembles in current operational use are likely to be too small, but how big do they need to be? In general, the number of samples that must be collected to estimate the distribution of a forecast variable depends on the form of the distribution, and in weather prediction there are many forms. Figure 1 shows a conceptual model of how the distribution of a forecast variable might evolve over time. The uncertainty in the initial conditions is relatively small, at least in comparison to the uncertainty at later times, and is commonly assumed to be Gaussian in the data assimilation system. Over time, the distribution will broaden and often develop asymmetric tails (left panel), for example due to constraints that quantities like humidity must be non-negative. As time advances and non-linear processes strongly influence the dynamics, the distribution may take a complex form with heavy tails representing higher probabilities of extreme events, or even multimodal distributions associated with preferred regimes (Fig. 1, centre panel). Eventually the forecast ensemble will lose all memory of the initial conditions, and the forecast distribution will converge to a typically broad, smooth, climatological distribution (right panel). Note that in the case of convective-scale forecasting in a limited-area model, the "climatological" distribution would represent the possible weather within the limited-area model when small-scale errors have saturated, subject to the range of synoptic-scale conditions represented in the driving global ensemble (Selz, 2019). This "climatological" distribution might be reached within a day or two (Hohenegger and Schär, 2007), but rather than being time-independent the distribution would continue to evolve on synoptic timescales.

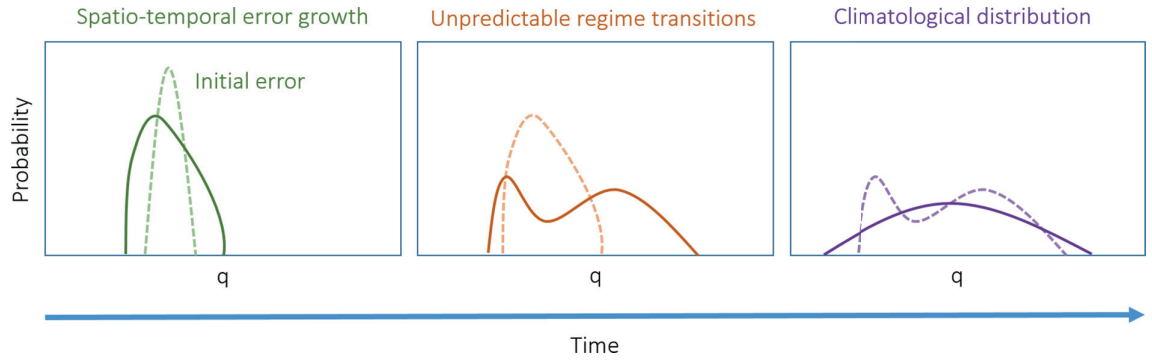

FIGURE 1 Conceptual model showing the time evolution of the distribution of a hypothetical forecast variable q. In each panel the dashed line represents the distribution at an earlier time, which evolves into the distribution shown by the solid line. In the centre and right panels, the dashed line is identical to the solid line in the previous panel. See text for further details [Colour figure can be viewed at wileyonlinelibrary.com]

In addition to changing over time, the shape of the forecast distribution will also be influenced by the forecasting problem being considered. It will depend on the weather regime (e.g., convective or clear), and also on the quantity being forecast. Variables such as precipitation cannot be negative and take skewed distributions, while aggregated quantities such as time- or area-averages may be more Gaussian than point values. It may be possible to estimate the ensemble mean accurately with a relatively small ensemble, but higher moments of the distribution, or the probability of extreme events, might require much larger ensembles (Leutbecher, 2019). The question of how big an ensemble should be is actually a set of many questions, and the cost of experimenting with large ensembles in NWP means that definitive answers are difficult to obtain.

Given the concern that the ensemble sizes that are computationally feasible may not be large enough to accurately represent the forecast uncertainty, a number of techniques have been proposed to increase the representativity of small ensembles. In global ensembles, perturbations based on singular vectors or bred vectors can be used to ensure that the ensemble captures the most rapidly-growing error modes (Palmer *et al.*, 1998; Toth and Kalnay, 1997). In limited area models, lateral boundary conditions can be chosen to ensure that the full spread of the global ensemble is represented (Montani *et al.*, 2011). Marsigli *et al.* (2014) show that lack of diversity in the global ensemble providing boundary conditions to a limited-area ensemble prediction system can be a major limitation when the global ensemble is small, or based on a single forecast model. The strength of this effect is likely to depend on the weather regime, as found by Keil *et al.* (2014) in their investigation of ensemble forecasts in a convection-permitting model. They found that the driving model is the dominant source of uncertainty when the synoptic forcing of convection is strong, but model physics perturbations that affect the triggering of convection are the main source of uncertainty when the forcing is weak.

In many forecast problems, particularly on the convective or sub seasonal to seasonal scales, predictions are made for averaged quantities, since much of the variability in the ensemble may be associated with rapidly varying weather systems that obscure predictable variations on larger scales (Toth and Buizza, 2019). If the small-scale variations are uncorrelated among the ensemble members, the variability will decrease as the averaging region increases. Finally, for probabilistic predictions of cumulus convection, it is often possible to increase the effective ensemble size by sampling the statistics within neighborhoods rather than for individual grid points (Ebert, 2009; Ben Bouallègue *et al.*, 2013; Hagelin *et al.*, 2017). This will be effective if the neighborhood is large enough that the convection at different grid points within the neighborhood is uncorrelated, and small enough that the statistical properties are homogeneous, that is not modified by factors such as large variations in orography or synoptic weather conditions. All of these methods can help offset the sampling errors due to small ensemble size, but their applicability depends on particular assumptions about the variability of the weather to be predicted.

The starting point for this article is the availability of an exceptional dataset, consisting of several 1,000-member convective-scale ensemble forecasts. Necker *et al.* (2020a) described the ensemble and evaluated the overall performance in comparison to an operational ensemble prediction system. They conducted an extensive analysis of the spatial and temporal covariance structures, motivated by two applications: ensemble data assimilation and localization, and estimation of ensemble sensitivities as a basis for evaluating the relative potential impact of different observations. The dataset from the large ensemble allowed them to estimate the errors in the covariances and sensitivity

diagnostics when restricted to smaller ensemble sizes. Building on this work, Necker *et al.* (2020b) evaluated a promising method for correcting for sampling error in these diagnostics when estimated from smaller ensembles. The present study uses the same set of forecasts, but is motivated by the use of ensembles to generate probabilistic forecast products. The 1,000-member ensemble dataset includes 14-hr forecasts, with a 3-km resolution, for eight different days that featured convective weather over Germany. Although the length and number of forecasts is limited, the large ensemble size provides an opportunity to characterize the forecast distributions with exceptional accuracy, and to address the question of how big an ensemble is required for a wide variety of forecast variables drawn from different distributions.

Our approach to investigating the 1,000-member ensemble has three main parts. First, we visually examine histograms of many forecast quantities, and attempt to classify them subjectively into a small number of characteristic distribution types. Second, we choose a representative forecast variable for each type, and examine the qualitative behavior of the distribution as function of neighborhood size, spatial averaging area, and forecast lead time. Finally, we compute the quantitative rate of convergence with increasing ensemble size of the mean, standard deviation and 95th percentile for each variable, or probability of exceeding a threshold in the case of precipitation. In the concluding section of the article, we will discuss how the results of the present study might generalize to other weather regimes and forecast systems, and identify some questions for future research.

2 | DATA AND METHODS

2.1 | The SCALE-RM ensemble

This study is based on 1,000-member ensemble forecasts from a limited area, convection-permitting model, and includes 10 ensemble forecasts for convective weather over Germany in the summer of 2016. The simulations were performed on the K-Computer at the RIKEN Center for Computational Science in Kobe, Japan.

The numerical model used for the forecasts is the open-source Scalable Computing for Advanced Library and Environment Regional Model (SCALE-RM: version 5.1.2; see Nishizawa *et al.*, 2015; Sato *et al.*, 2015; Nishizawa and Kitamura, 2018). The simulations use the Tomita (2008) single-moment bulk microphysics scheme, the Mellor–Yamada–Nakanishi–Niino 2.5 closure boundary-layer scheme (Nakanishi and Niino, 2004), the Model Simulation Radiation Transfer code for the representation of radiative fluxes (Sekiguchi and Nakajima, 2008), and a Beljaars-type surface model (Beljaars d Holtslag, 1991) for the computation of soil variables and surface fluxes. The horizontal resolution is 3 km, and the domain covers 350×250 grid points with 30 vertical levels (Fig. 2).

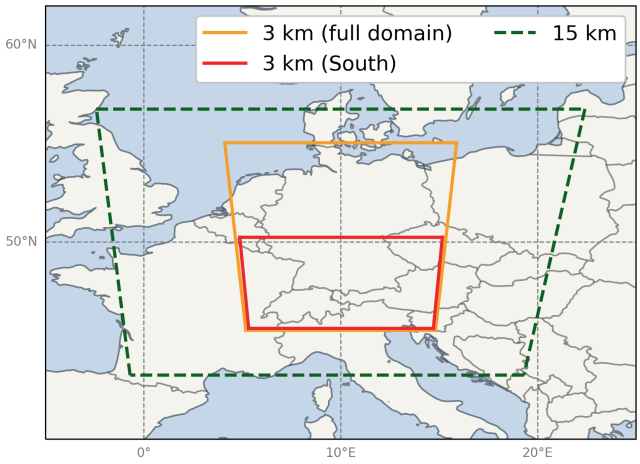

FIGURE 2 Model domains for the Scalable Computing for Advanced Library and Environment Regional Model (SCALE-RM) simulations: 15-km resolution model domain used for data assimilation cycling and boundary conditions (dashed line); full domain for the 3-km forecasts (solid line, larger region), and the Southern subdomain used for analysis (solid line, smaller region) [Colour figure can be viewed at wileyonlinelibrary.com]

Initial and boundary conditions are obtained from forecasts using a 1,000-member SCALE-RM ensemble with 15-km resolution, which in turn uses boundary conditions from the 20-member Global Ensemble Forecast System (GEFS) of the National Center for Environmental Prediction (NCEP). The 15-km resolution ensemble uses a Local Ensemble Transform Kalman Filter system (SCALE-LETKF: Lien *et al.*, 2017) to assimilate conventional data with a 3-hr assimilation window. The ensemble DA was initialized from a previous 15-km, 1,000-member DA experiment over the same domain, spun up for a week. Boundary conditions for the 15-km forecast were prepared every 6 hr, using a combination of randomly generated perturbations and the 20-member GEFS analyses. The boundary perturbations were generated by taking the difference between two randomly chosen atmospheric states corresponding to the same season and time of day, then scaling the amplitude by a factor of 0.1. Every 6 hr, each perturbation was updated using the difference between the atmospheric states 6 hr after the states used for the previous perturbation, ensuring that the perturbations evolve smoothly. The initial conditions for the 3-km ensemble were obtained by interpolating the corresponding fields from the members of the 15-km analysis ensemble, and boundary conditions were provided hourly by a parallel 15-km ensemble forecast on the outer domain. Ensemble

forecasts were initialized every 12 hr (at 0000, 1200 UTC) over an eight-day period from 0000 UTC, 29 May to 1200 UTC, 7 June (excluding 3 and 4 June), for a total of 16 forecasts. The forecasts were integrated out to a lead time of 14 hr. Further details of the ensemble set-up are provided by Necker *et al.* (2020a,b).

2.2 | Synoptic situation

The period from 29 May to 7 June 2016 was characterized by a quasi-stationary weather pattern over central Europe. An upper-level trough over western Europe was accompanied by a shallow surface low in the first part of the simulation period, followed by a weak pressure-gradient synoptic pattern (Fig. 3). The mid-level winds started warm and moist southerly to easterly in the first half of the period and backed to northeasterly at later times.

Throughout the simulation, the environment was highly unstable, through a combination of synoptic forcing and surface heating, with the latter process increasing in importance throughout the period. Cumulus convection was most intense in the first few days and the motion of the convective cells was slow. Later in the simulation period, convection followed a clear diurnal cycle, with a peak late in the afternoon. Because of the intense, extensive and recurring convection, this period has been the subject of many investigations (e.g. Piper *et al.*, 2016; Necker *et al.*, 2018; Keil *et al.*, 2019; Bachmann *et al.*, 2020; Scheck *et al.*, 2020).

For this study we focus on forecasts started at 1200 UTC on three selected days: 31 May, 29 May and 5 June 2016. The figures will show examples from 31 May 2016, since this day featured a combination of patchy convective showers and more widespread precipitation regions (Fig. 4a–c). As a result, this day showed the most varied and complex distributions of forecast variables. Regarding the other two days, 29 May exhibited more intense and organized convective activity over the domain, and 5 June much weaker large-scale forcing and more scattered convection, mostly over orography. A preliminary investigation of the forecast probability distributions on these other days showed a similar range of behaviors to those of the 31 May forecast, and no additional new features. Similarly, results will only be shown for the southern half of the model domain, since the northern region did not provide any additional behaviors or insight.

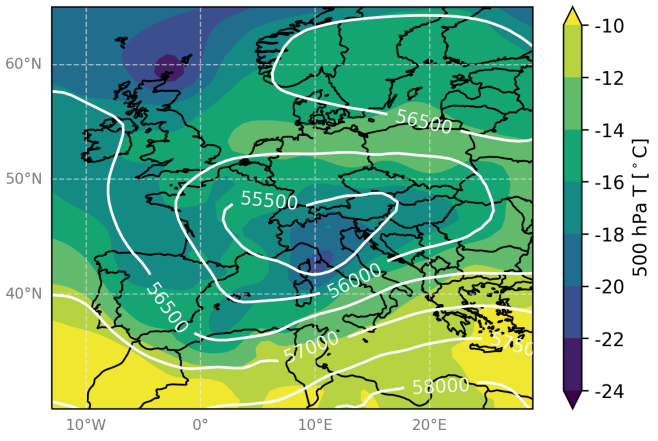

FIGURE 3 ECMWF ERA-Interim reanalysis of 500 hPa geopotential ($m^2 \cdot s^{-2}$, white contours) and temperature (°C, shaded) for 1800 UTC 31 May 2016 [Colour figure can be viewed at wileyonlinelibrary.com]

2.3 | Forecast variables

A 1,000-member ensemble generates vast amounts of data, which is challenging for both storage and movement of the datasets. For the present investigation we extracted a subset of the archived data that covered many of the variables most relevant for convective-scale weather prediction. The selected variables were three-dimensional fields of temperature, horizontal and vertical wind velocities (u, v, w), specific humidity, relative humidity, hydrometeor mixing ratio and radar reflectivity, on selected pressure levels and two-dimensional fields of mean sea-level pressure, surface precipitation rate, surface net longwave radiation flux, surface net shortwave radiation flux, surface downward longwave radiation flux, surface downward shortwave radiation flux, top of atmosphere (TOA) net longwave radiation flux and shortwave radiation flux. The data was stored at hourly intervals for each of the ten 14-hr forecasts.

Even this dataset was too large to be analyzed fully, and we have further restricted our analysis to temperature, horizontal and vertical wind velocity, reflectivity, and specific humidity at 500 hPa, as well as surface precipitation rate and mean sea level pressure. Regarding specific humidity, it soon became apparent that an important factor influencing the distributions was the distinction between cloudy (saturated) regions and unsaturated air. To make this distinction more obvious, we have generally plotted specific saturation deficit (q_{def}), instead of specific humidity. This quantity is defined as the difference between the saturation water vapor mixing ratio at the grid point temperature and the actual mixing ratio and is related to relative humidity:

$$q_{def} = q_{sat} - q = q\,(RH^{-1} - 1), \tag{1}$$

where $RH = q/q_{\mathrm{sat}}$ is the relative humidity. An example of the ensemble mean q_{def} field is shown in Fig. 4c.

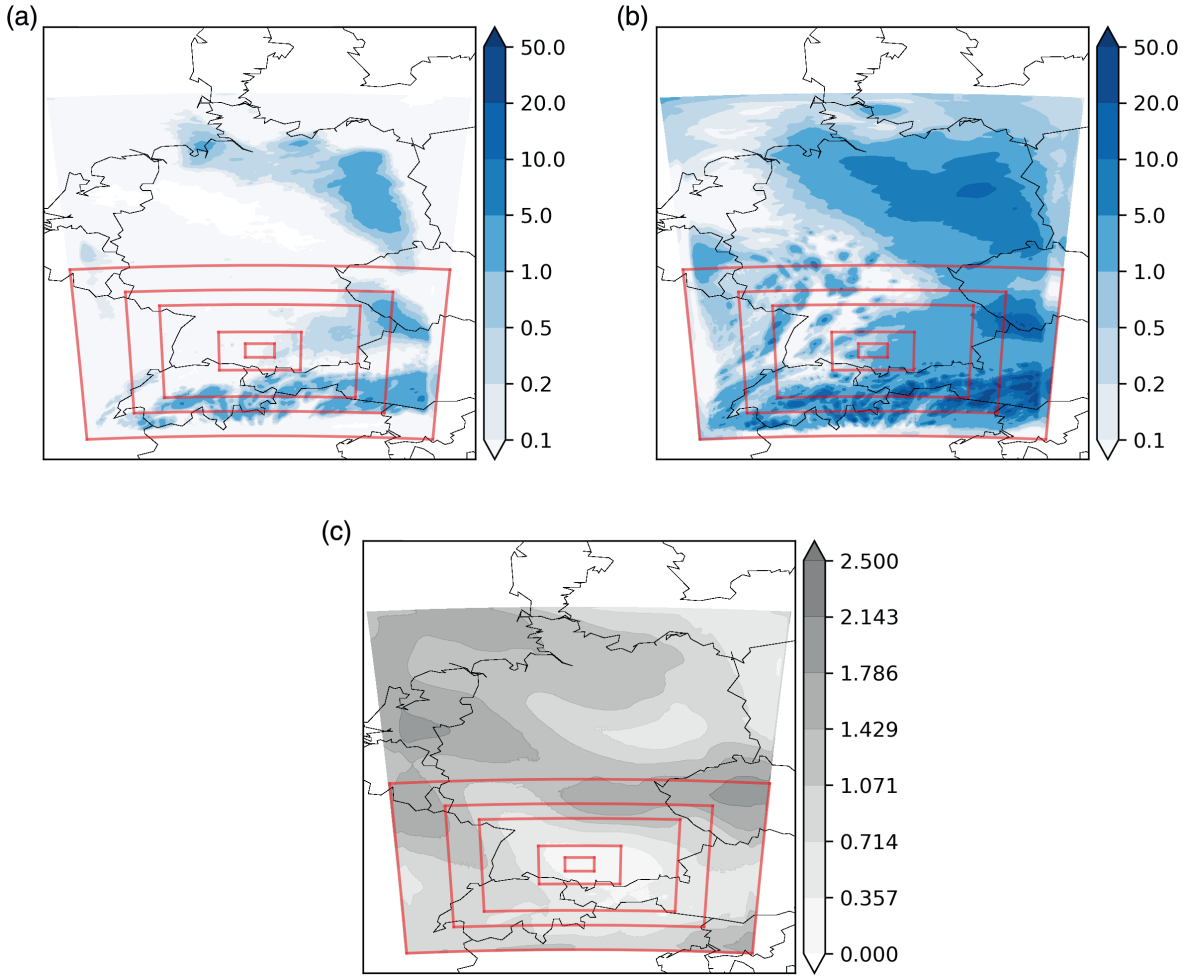

FIGURE 4 (a) SCALE-RM ensemble mean hourly precipitation (mm), plus 6 hr lead time; (b) 14 hours accumulated precipitation (mm); and (c) specific saturation deficit(q_{def}, g·kg^{-1}) at 500 hPa, 6 hr lead time (for the forecast started at 1200 UTC on 31 May 2016). The rectangles in (a–c) are the subdomains used in the analysis of spatial-scale dependencies (see Section 2.4) [Colour figure can be viewed at wileyonlinelibrary.com]

2.4 | Spatial scales and time evolution

Histograms were plotted for the selected variables, showing the variability across the 1,000 ensemble members at individual grid points and times. Each plot was constructed using 50 bins, and shows the number of occurrences that fall within each bin, divided by the total number of occurrences and the bin width. Histograms were also produced for different spatial regions, defined in Figure 4. The meridional widths of the chosen subregions are 3, 45, 120, 285, 375 and 525 km. The largest region consisted of the southern half of the model domain, with size 525 × 750 km. The other subregions were chosen to have the same shape, for example 45 × 63 km (15 × 21 grid points), except for the 3-km region, which corresponds to a single grid point. Distributions of variables were examined for these six subdomains, centered over the middle point of the largest subdomain (Fig. 4) for a selected time (31 May 2016, 1800 UTC, at 6 hr of lead time). The largest sub region extends to the boundaries of the model domain. This raises the possibility of a penetration of boundary condition influence into the analysis region, most dramatically taking the form of a spin-up of cloud and precipitation variables near inflow boundaries. Figure 3 suggests that this will not be a major influence in the present case study, since the synoptic flow is weak everywhere. The boundaries might be a factor in the low precipitation totals on the eastern edge of the domain (Fig. 4b), but this is only relevant for the largest subdomain and even there affects only a small fraction of the area.

Histograms for spatial regions were produced in two ways. In the neighborhood method, the individual grid points in the chosen region are treated as independent ensemble members. This increases the effective size of the ensemble, but only provides new information if the different grid points are uncorrelated. The second method plots

histograms of area-averaged quantities. In this case, the spatial variability is averaged out if it is uncorrelated in space, leading to less variability across the 1,000 members and a less uncertain forecast.

Finally, the time evolution of the distributions was examined. Results were computed using the neighborhood method for a fixed region size (45-km width), and for lead times from the first time step (which will be referred to as the initial time) to 14 hours. Histograms were then compared to the conceptual model of the evolution of a forecast distribution in Figure 1.

2.5 | Sampling error and measuring convergence

Estimates of forecast quantities constructed from a small ensemble will suffer from sampling error, but should converge to an accurate value as ensemble size increases. To provide a quantitative measure of this convergence, estimates were made using a range of ensemble sizes, subsampled from the 1,000-member ensemble. In general the members of the sub ensembles were selected randomly, but for the set of estimates based on "selected members" we constrained the selection to draw equal numbers of members using boundary conditions from each of the 20 members of the GEFS ensemble. If the synoptic environment, which is represented by the spread in boundary conditions from the global ensemble, is a main source of uncertainty, it may be possible to capture the most important characteristics of the distribution in a small ensemble, as long as this variability is explicitly included. The selected member ensembles were defined by Necker *et al.* (2020a) for ensemble sizes of $N = 20$, 40, 80, 160, 320, 600 and 800.

Confidence intervals for the estimates were constructed as follows. For each ensemble size, 10,000 test ensembles were created by bootstrapping with replacement. The forecast parameter (e.g., ensemble mean) was computed from each test ensemble to create a distribution of estimates. The 2.5th and 97.5th percentiles of this distribution then define the 95% confidence interval.

3 | RESULTS AND DISCUSSION

3.1 | Classification of distributions

The first goal of this investigation is to inspect histograms of the various forecast quantities for different regions and times, in order to identify the characteristic types of distribution produced by the ensemble. Each forecast variable was found to have a typical shape, and these shapes could be classified into three broad categories: quasi-normal distributions, highly skewed distributions, and mixtures with two or occasionally three peaks. Table 1 shows which variables are assigned to each category. The remainder of this subsection presents examples of each category to illustrate their distinctive properties.

TABLE 1 Classification of variables. All three-dimensional variables are extracted at 500 hPa

Category	Variable
Quasi-normal	Temperature
	Horizontal wind velocity
	Vertical wind velocity
	Mean sea level pressure
Highly skewed	Precipitation
	Reflectivity
	Specific humidity
Mixture	Specific saturation deficit
	Relative humidity

A distribution is classified as quasi-normal if it is unimodal, with a relatively small skew. In most cases, these distributions are fitted well by a Gaussian function. Variables with this distribution shape include temperature (see Fig. 5a), all wind components at 500 hPa, and mean sea-level pressure. The quasi-normal shape was found for all neighborhood widths, averaging regions, and forecast lead times. Note that this subjective description does not take into account outlier values, such as the temperature or vertical velocity anomalies at the core of a convective updraft, which are very rare (around 0.1% of grid points) in the case considered here.

Variables showing highly skewed distributions include precipitation and reflectivity. These quantities are both related to hydrometeor content and hence bounded by zero. The example precipitation distribution shown in Figure 5b is closer to lognormal than normal in shape. Note that the full distribution of precipitation rates includes a point mass at zero representing members that have zero precipitation at this location, and would be best described as a mixture that resembles a combination of a lognormal distribution and delta function at zero. We have plotted the histogram with a logarithmic x-axis to make the relation to a lognormal shape more visible, but with the disadvantage that the zero precipitation values cannot be plotted. Instead, a numerical value for R, the fraction of ensemble members with non-zero precipitation is given (precipitation rates greater than 10^{-10} mm·hr^{-1} are considered to be non-zero). In Figure 5b, $R = 0.39$ indicates that 39% of the ensemble members have non-zero

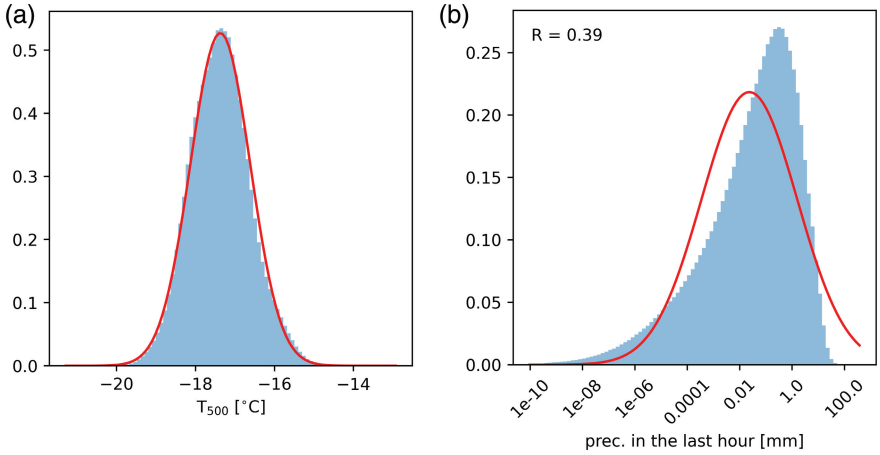

FIGURE 5 Histograms of: (a) temperature (°C) at 500 hPa; and (b) precipitation [mm] in the last hour on 31 May at 1800 UTC (forecast lead time 6 hr) in a 375 km wide neighborhood. Note the logarithmic axis for precipitation in (b), and the value R denoting the fraction of members with non-zero precipitation. A Gaussian function with the same mean and standard deviation (a) and a lognormal function (b) are shown for comparison (solid lines) [Colour figure can be viewed at wileyonlinelibrary.com]

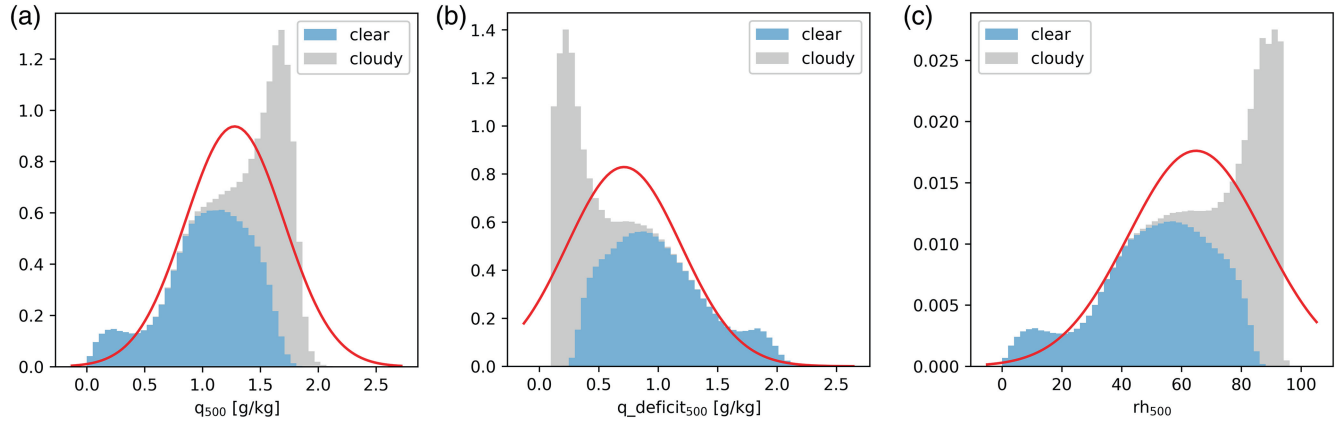

FIGURE 6 Histograms of: (a) specific humidity ($g \cdot kg^{-1}$); (b) specific saturation deficit ($g \cdot kg^{-1}$); and (c) relative humidity (%) at 500 hPa in the southern region with a width of 375 km on 31 May, 1800 UTC (forecast lead time 6 hr). Grey and blue colors indicate contributions from cloudy and clear grid points at 500 hPa, respectively. The criterion to distinguish cloudy grid points is a simulated radar reflectivity higher than −19 dBZ at 500 hPa. A Gaussian function with the same mean and standard deviation is shown for comparison (solid lines) [Colour figure can be viewed at wileyonlinelibrary.com]

precipitation, while the remaining 61% have zero values and are not seen on the plot.

The last group, mixture distributions, includes the specific saturation deficit, q_{def}, and other humidity variables. Figure 6 compares the distributions of specific humidity to relative humidity and q_{def}, with the latter two variables showing clearly how the moisture distribution is bounded by the saturation humidity. The figure also shows that the complex distribution shape arises as a mixture of distinct distributions in cloudy and clear regions. In this particular example, there is even a third peak in the humidity distribution, since this example is computed for a large neighborhood (width 375 km), that includes not only the mixture of cloudy and clear grid points in the rainy region in southern Germany, but also part of the dry band across the central part of the country (Fig. 4c).

The three types of histogram presented here represent very broad categories, identified subjectively, rather than parametric distributions that can be quantitatively fitted. They are useful to illustrate the different dependencies of distribution shape on aggregation over regions of different sizes, and on forecast lead time. In the following subsections, we will illustrate these dependencies using one example variable from each of the three categories. Details of the distributions will depend on the specific weather situation experienced at that place and time, and in general, it would be interesting to consider other locations, times and forecast variables. However, the limited period of our forecast dataset means that even if we considered all possible distributions, the results would still not be representative of the range of behaviors encountered by a regional ensemble forecasting system. Instead, we have chosen to analyze a small number of examples in depth, and in Section 4, we are careful to distinguish results that might be generally applicable from those that may be specific to this location, time, and weather situation.

3.2 | Scale dependency

We now examine qualitatively how the forecast distributions are represented in the 1,000-member ensemble. We consider distributions for three variables, representative of the three categories when computed for single grid points, and apply two methods that use spatial information to improve the representation. As described earlier, the neighborhood method treats grid points within a spatial region as alternative, but statistically equivalent, realizations of the ensemble at the target grid point, giving a larger effective sample size. This method is expected to work only to the extent that the grid points are uncorrelated. If this assumption is true, the histogram will not change form as neighborhood size increases, but will become smoother and better defined, as sampling error is reduced. The averaging method considers the distribution of an area-averaged quantity. Averaging will remove small-scale variability, but again only to the extent that the grid points are uncorrelated in space. If this is true, the averaging operation will be a sum of independent random variables, and the Central Limit Theorem ensures that the histogram will converge to a Gaussian distribution, with variance that decreases as the size of the averaging region increases (Bonamente, 2017).

For each method, the scale dependency is illustrated by analyzing distributions of variables on 31 May 2016 at 1800 UTC for the six concentric subdomains shown in Figure 3. This location was chosen mainly for convenience, since it lies in the centre of the southern domain, allowing a large range of region sizes to be used. However, this should also be an interesting region, since it includes a variety of precipitation types and intensities.

3.2.1 | Distributions over neighborhoods

Figure 7 shows distributions of example forecast variables from each of the three categories for different neighborhood sizes. The three rows show histograms of temperature at 500 hPa, 1-hr accumulated precipitation, and specific saturation deficit at 500 hPa. The first column shows the distribution across the ensemble at a single grid point. Here, 1,000 members are sufficient to produce a smooth distribution for temperature, but not for the two other variables. This is particularly true for precipitation, where only 21% of members have precipitation at all.

The second column in Figure 7 shows distributions for a neighborhood with width 45 km. As detailed in Section 2.4, this neighborhood contains $15 \times 21 = 315$ grid points, giving an effective ensemble size of 315,000 if the grid points are truly independent. In reality, the effective resolution of the model is likely 5–6 grid points, leading to correlations between nearby grid points. In this case the number of independent grid points in each direction will be correspondingly smaller, reducing the effective

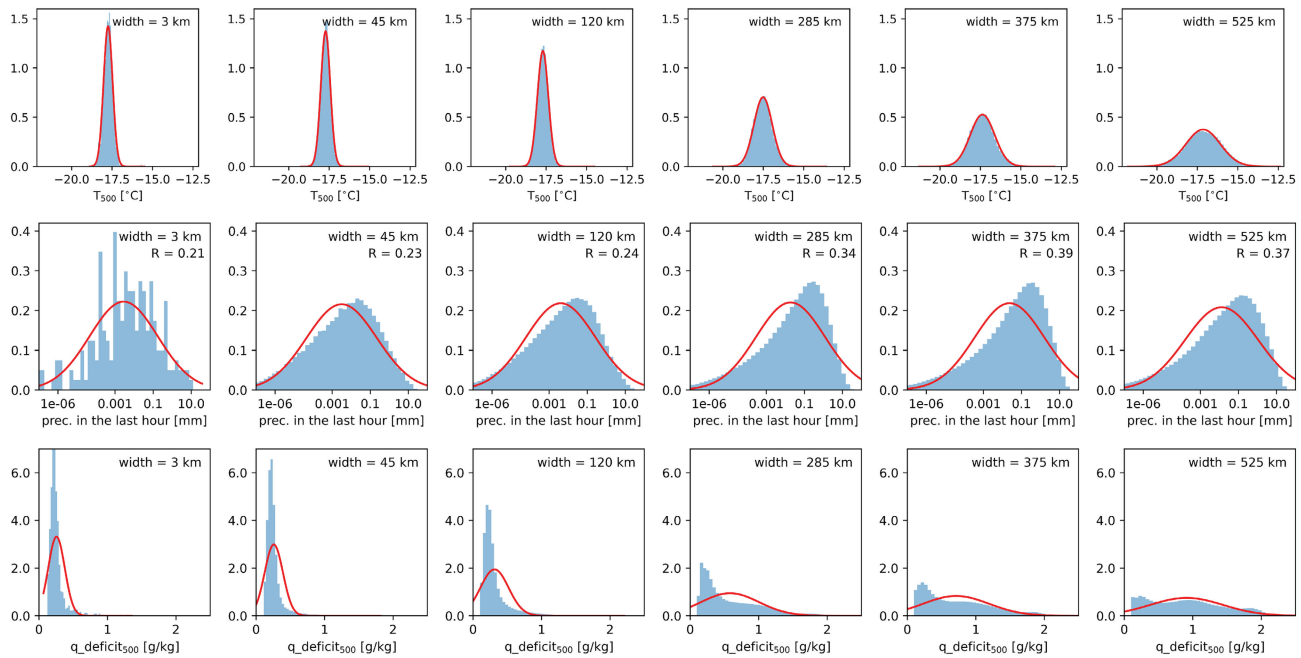

FIGURE 7 Histograms of temperature at 500 hPa (first row), precipitation (second row) and specific saturation deficit at 500 hPa (third row) on 31 May 2016 at 1800 UTC (forecast lead time 6 hr). The histograms are computed for neighborhoods with widths of 3, 45, 120, 285, 375 and 525 km (columns from left to right). A Gaussian or log-normal distribution (for precipitation) with the same mean and standard deviation is shown for comparison (solid lines) [Colour figure can be viewed at wileyonlinelibrary.com]

ensemble size by a factor of 25–36 to about 10,000. In the case of larger convective systems, the effective ensemble size would be even smaller, but still might be substantially larger than the 1,000 members available at a single grid point, Indeed the histograms for the 45-km region are now smooth for all three variables. While the distributions are better defined for the 45-km neighborhood than for a single grid point, there is little evidence that the shapes of the underlying distributions have changed, which would be evidence of spatial correlations on larger scales. Instead, this behavior is consistent with the hypothesis that the grid points included in a small neighborhood are to some extent statistically independent and can be considered as additional ensemble members. In Section 3.4, we will attempt to provide a quantitative estimate of the effective ensemble size.

The relative independence of the forecast distributions on neighborhood size is no longer true when neighborhoods larger than about 100 km are chosen, as seen in the three columns on the right of Figure 7. The temperature distribution broadens with increasing neighborhood size. Since the spatial variation in the temperature field is mostly at synoptic scales, the distribution is quite narrow for small neighborhoods, but broadens as more large-scale variability is included in the larger region. On the other hand, the precipitation distribution narrows with increasing neighborhood size and the fraction R of members with precipitation increases. These effects are also likely to be associated with variability on large spatial scales, since the larger neighborhoods will include more of the broad regions of intense precipitation in the eastern half of the model domain (Figure 4a,b). The humidity distribution shows the effects of large-scale variability even more dramatically, with the distribution becoming very broad, and eventually showing a second and even a third maximum, as different geographical regions are included in the neighborhood.

The results shown in Figure 7 are likely to be specific to this case, time, and even to the selected grid point. However, some aspects of the qualitative behavior are likely to apply more generally. Variables such as precipitation and humidity benefit from the neighborhood method since their spatial variability includes a significant component that is small-scale and random. On the other hand, temperature variations in the troposphere are weak on smaller length scales and the increase in effective ensemble size due to the neighborhood brings no new information, leaving the distribution similar to the single grid point version. For all variables, a sufficiently large neighborhood will include synoptic variability, and the neighborhood distribution will no longer be representative of the distribution for the target grid point. For the variables shown here, the neighborhood must be smaller than about 100 km to avoid these effects, but the threshold is likely to be different for other weather situations, or other variables such as 2-m temperature, where the small-scale variability is strongly influenced by local factors such as land surface properties.

3.2.2 | Distributions of spatially averaged quantities

Histograms for the three example forecast variables, averaged over areas of different size, are shown in Figure 8. The left column shows the same single grid point data as the left column of Figure 7, but the appearance has changed due to different scales on the x- and y-axes.

The convergence of the distribution to a Gaussian form with larger averaging area is clearly seen for q_{def} in the bottom row of Figure 8. By an area width of 285 km, the histogram is largely symmetrical, and becomes narrower for larger averaging areas. The approach to a narrowing Gaussian distribution is seen in the second row for the logarithm of precipitation, coinciding with 100% of members having non-zero precipitation rates in the area average. This convergence to a lognormal distribution would be expected from the Central Limit Theorem for a product of independent random variables.

The distribution of temperature appears Gaussian in the 1,000-member ensemble even for a single grid point, but there is no decrease in width as the averaging area increases. Consistent with the results of the neighborhood analysis, this is likely related to the variability of the temperature being dominated by the synoptic scales. Note however, that this could act in different ways. The simplest possibility is that the temperature variability might be perfectly correlated across the subsynoptic averaging regions, so that there is no small-scale variability to average out. On the other hand, it might be that any decrease in small-scale variability due to averaging is offset by the inclusion of increasing amounts of large-scale variability as the region size increases.

3.3 | Time evolution

In this section we examine the time evolution of the forecast distributions. The conceptual model that was presented in the introduction (Fig. 1) suggested that we might expect the histogram to evolve from a simple form given by the data assimilation (perhaps Gaussian), to a more complex, even multimodal, shape. Eventually the distribution would converge to a smooth "climatological" form, that in the case of a convective-scale forecast, would be determined by the large-scale weather pattern when predictability on the convective scale is lost. As an example of this process, Figure 9 shows the time evolution of

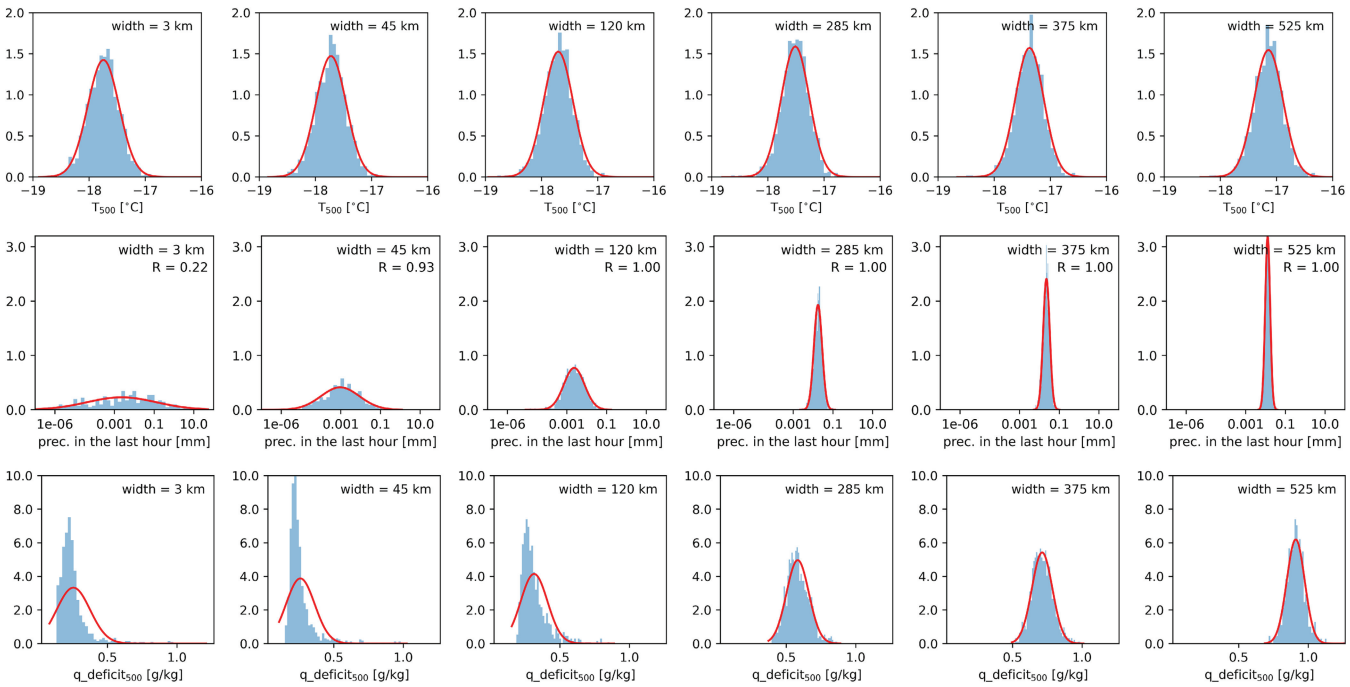

FIGURE 8 As in Figure 7, but computed by averaging over subdomains of width 3 (single gridpoint), 45, 120, 285, 375 and 525 km [Colour figure can be viewed at wileyonlinelibrary.com]

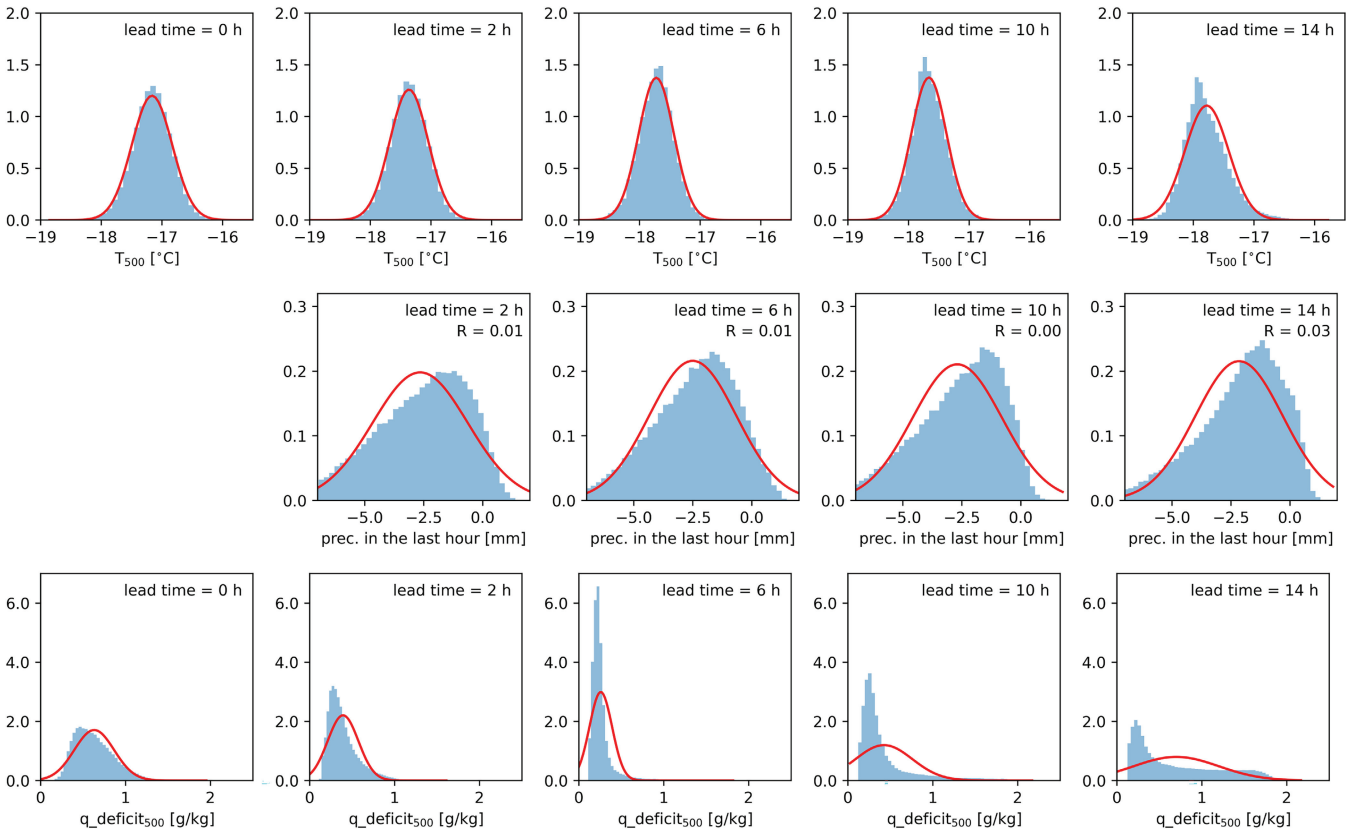

FIGURE 9 Histograms of temperature (top row), precipitation (middle row) and specific saturation deficit (bottom row) on 31 May 2016 in southern Germany, for a neighborhood of width 45 km. Forecast lead time increases from left to right from the first time step (1200 UTC) to 14 hr (0200 UTC, 1 June). A Gaussian or log-normal distribution function (for precipitation) with the same mean and standard deviation is shown for comparison (solid lines) [Colour figure can be viewed at wileyonlinelibrary.com]

the three selected variables in terms of histograms over 45-km neighborhoods. This neighborhood provides a large enough effective ensemble size to visualize the distributions, but not so large as to include significant variability on large spatial scales.

The top row of Figure 9 shows that the temperature at 500 hPa is gradually cooling over the 14 hr of the forecast. The distribution starts close to Gaussian, and remains so for several hours, with little change in width. Eventually the histogram develops a significant skew. Figure 3 shows that the temperature field is dominated by a synoptic-scale gradient, with colder air to the south. The cooling trend shows that colder air is being advected north on average, but this non-linear process proceeds at different rates in different ensemble members and does not preserve the Gaussian distribution. Overall, the behavior is consistent with the first two stages of the conceptual model, but the evolution is occurring on a relatively slow, synoptic time-scale.

The evolution of the distribution of precipitation is shown in the second row of Figure 9. Since precipitation is accumulated over the previous hour, it is not available at the initial time. Starting from a lead time of two hours, the histogram changes only slowly as the synoptic environment evolves. Since the predictability of convective clouds is very short, it is possible that the precipitation distribution moves through all stages of the conceptual model within two hours, arriving at a "climatological" distribution determined by the large-scale flow. Unfortunately, it is not possible to test this interpretation with the available data.

The humidity distribution starts fairly close to Gaussian at the analysis time, but narrows and becomes increasingly asymmetric over time. At 6 hr lead time, the histogram is concentrated near the zero bound of saturation deficit, indicating that the region has become cloudy in almost all ensemble members. By the end of the forecast at 14 hr, the histogram shows both a cloudy and a clear-sky peak. It is not known whether the distribution will continue to evolve in time, but this form would be expected for the "climatological" distribution, when predictability of the locations of the convective clouds is lost. At this point, the overall humidity is set by the large-scale conditions, but there is no skill in predicting whether a particular grid point will be cloudy or clear.

Overall the behavior of the distributions is not inconsistent with the conceptual model of Figure 1, but it is not possible to clearly identify all of the stages for any of the forecast variables shown here. At the very least a higher time resolution would be needed to see the impact of the initial loss of predictability of the convection, while a longer simulation time might be needed before the distributions could be regarded as climatological, even in the sense defined for Figure 1.

3.4 | Convergence of ensemble predictions

The quantitative convergence of probabilistic predictions with ensemble size will be explored by considering distributions of temperature, precipitation and humidity obtained in three ways: first for grid point values, second for a 45-km neighborhood, and finally for grid point values obtained from a "selected member" ensemble constrained to sample all 20 large-scale boundary conditions equally (see Section 2.5). For temperature and humidity, we examine convergence of the ensemble mean, the standard deviation, and as a representative of more extreme events, the 95th percentile. For precipitation, the standard deviation and 95th percentile of the distribution are less useful because of the large number of members with zero precipitation. Instead, we consider the probability of precipitation exceeding two thresholds: 0.1 and 5 $mm \cdot hr^{-1}$. The first threshold is a very modest amount of rain, whereas the second threshold corresponds to a relatively rare event in the ensemble.

As explained in Section 2.5, a distribution of estimates for each forecast quantity was constructed by bootstrapping. Figure 10 shows the bootstrap sample median and confidence intervals for each estimated quantity as function of ensemble size. Starting with the grid point estimates for the quasi-normal variable, temperature, the top left panel shows that the estimate of the mean for all ensemble sizes is unbiased and the confidence interval narrows as ensemble size increases. In contrast, estimates of the standard deviation of temperature (second column), and even more so the 95th percentile (third column), are biased low for small ensemble sizes, with the median value below that of the 1,000-member reference (shown in white). The rate of decrease in the width of the confidence interval for smaller ensemble sizes is irregular, although it becomes smoother for larger ensemble sizes. For all three quantities, the neighborhood method produces slightly narrower confidence intervals than the single grid point values, consistent with a larger effective ensemble size.

Similar conclusions can be drawn for the precipitation and humidity estimates in the second and third rows of Figure 10, with irregular behavior for small ensemble sizes, but tending towards a smoother convergence behavior for larger ensembles. The bias of the estimate is particularly pronounced for the higher precipitation threshold of 5 mm h^{-1}. The median probability is zero for ensemble sizes of 40 or less, indicating that this precipitation

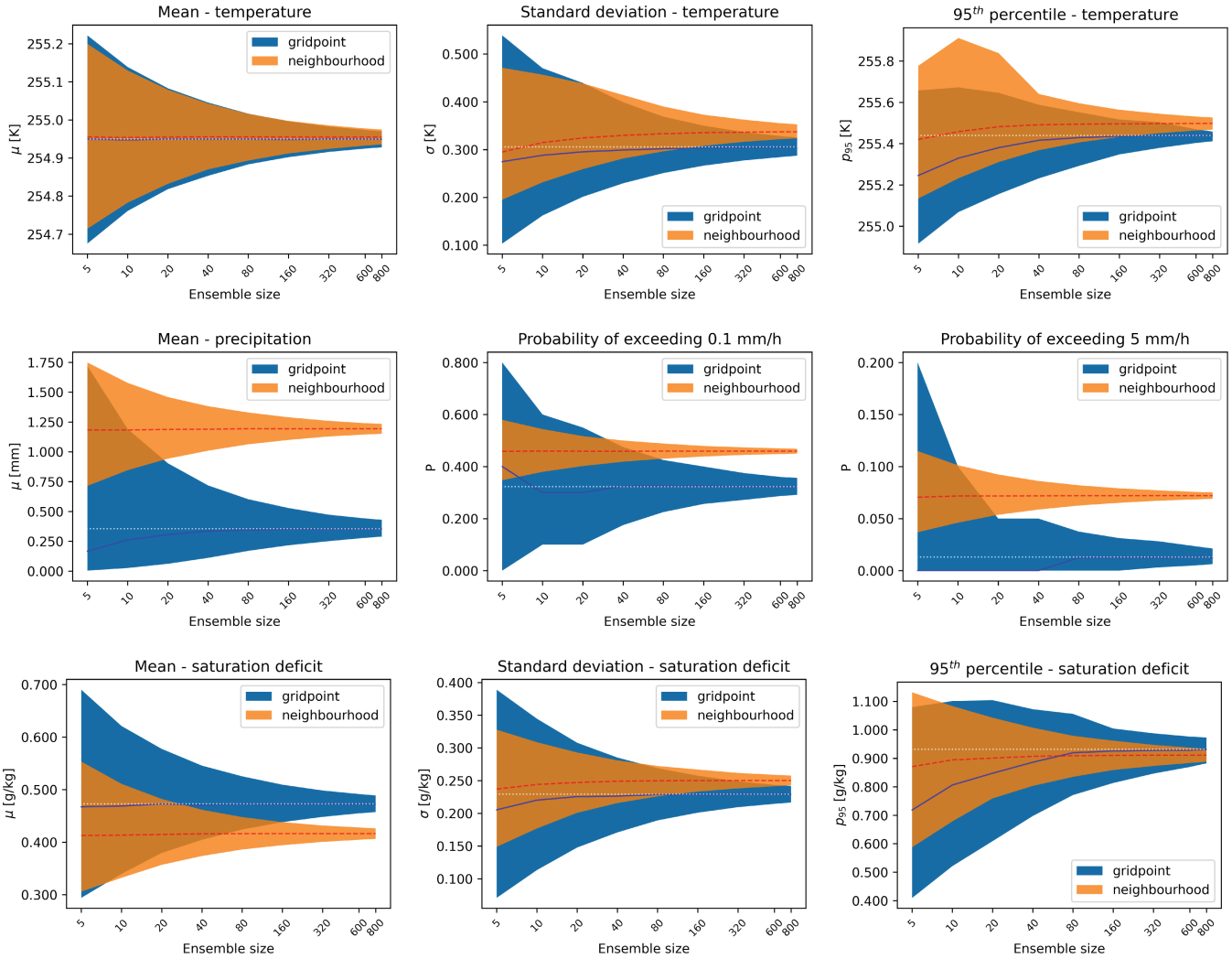

FIGURE 10 Mean, standard deviation and 95th percentile of 500 hPa temperature (top row) and saturation deficit (bottom row). Middle row shows mean hourly precipitation and probability of precipitation exceeding thresholds of 0.1 and 5.0 mm·h^{-1} Forecast quantities are computed for 31 May at 1800 UTC (forecast lead time 6 hr). The bands show the 95% confidence interval determined for the 10,000 bootstrapped samples (bounded by 2.5% and 97.5% quantiles), while the solid lines show the respective median values. The dashed white horizontal line indicates the median of the distribution computed using all 1000 members. Grid point: single grid point forecast, neighborhood: 45-km neighborhood forecast [Colour figure can be viewed at wileyonlinelibrary.com]

rate did not occur at all in more than half of the bootstrap ensembles. Comparing the grid point results to the neighborhood estimates, it is clear that the benefit for narrowing the confidence intervals is greater for precipitation than for temperature. However, the neighborhood estimates of ensemble mean and other quantities converge to a value that is different from the large ensemble limit computed for a single grid point. The difference is particularly prominent for precipitation. As shown in Fig. 4, the location under consideration is on the edge of a region of frequent precipitation, so that the neighborhood includes locations with higher probabilities of more intense rain. As a result, the variability in the neighborhood is not an accurate substitute for the ensemble variability at a single grid point in this case.

Finally, we consider the performance of the selected member ensembles, which are expected to increase the spread of small ensembles by always including all large-scale boundary conditions. The results of these experiments have not been plotted in Figures 10 and 11, because they were almost identical to those of the randomly chosen ensembles. No impact on ensemble spread was found. This result is consistent with the analysis of Necker *et al.* (2020a), who showed that the inclusion of different large-scale boundary conditions in the ensemble had only a small effect on the variance spectrum during

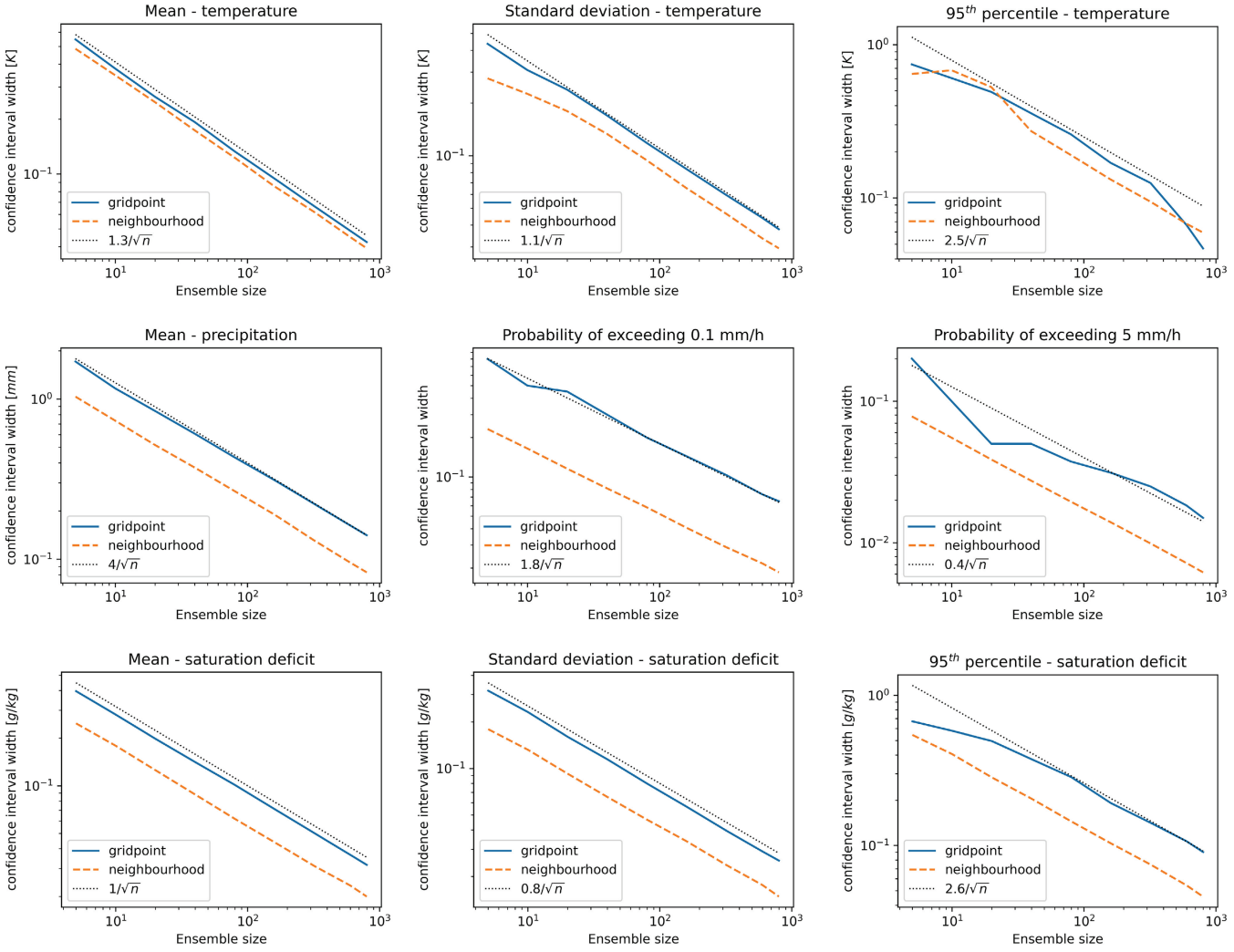

FIGURE 11 Width of the 95% confidence intervals, between 2.5% and 97.5% quantiles, based on 10,000 bootstrap samples, for the forecast variables shown in Figure 10. Dashed lines show reference curves with slope $N^{-1/2}$, fitted by eye [Colour figure can be viewed at wileyonlinelibrary.com]

the initial spin-up phase of the forecast, and no significant impact at later times. A possible explanation for the lack of impact is that, for the weather situation considered here, the variability in the ensemble is dominated by local processes (Keil *et al.*, 2014). In this case, a larger number of case studies with a greater variety of weather will be required to assess the value of a selected member ensemble. But it might also be that the 20 member GEFS ensemble does not adequately represent the large-scale uncertainty in this case, and the addition of random perturbations fails to compensate. In this case, it would be beneficial to use boundary conditions from a larger or even multimodel global ensemble (Marsigli *et al.*, 2014).

We now examine the rate of convergence of the forecast estimates by plotting the width of the confidence interval as a function of the ensemble size N. If the ensemble size is large enough, and the underlying distribution is well behaved (e.g., has finite moments), the Central Limit Theorem (CLT) states that for a large number of independent and identically distributed random variables, the sampling distribution of the normalized sum will tend towards a normal distribution without dependence upon the underlying distribution's shape (Dekking *et al.*, 2005). The standard error of the mean of this sampling distribution will then be proportional to $N^{-1/2}$. This normality of the sampling distribution can also be extended to a wide range of other statistics, including those of the standard deviation and of quantiles (Walker, 1968). The width of the 95% confidence interval is a multiple of the sampling distribution's standard error, and can also be expected to converge as $N^{-1/2}$. However, this behavior is only expected in the limit of large ensemble size, and it is not certain whether it will be observed for the meteorological distributions and ensemble sizes available here.

The first column of Figure 11 shows that for all three forecast variables, for random and neighborhood ensembles, the width of the confidence interval for the ensemble mean decreases proportional to $N^{-1/2}$. It is quite striking that the convergence follows the asymptotic law even for ensemble sizes of less than 10 members. The confidence intervals for the neighborhood estimates are always narrower than for the estimates based on a single grid point, consistent with a larger effective ensemble size. The difference is substantial for precipitation and humidity. By shifting the neighborhood curve to the right, we can estimate that the effective increase in ensemble size is approximately a factor of 3. Since the neighborhood contains 315 grid points, this shows that there is substantial correlation between the variability at different locations within the region. Interestingly, this ratio is consistent with an effective correlation length of approximately $(315/3)^{1/2} \approx 10$ grid points. In a numerical model, parameterized and numerical diffusion cause nearby grid points to be correlated, so that only features larger than 5-6 times the grid length are accurately resolved. The correlation length estimated here is about twice this size, suggesting that the effective ensemble size may be limited by physical interactions in space, rather than by the effective model resolution, as proposed in Section 3.2.

For the estimates of standard deviation of temperature and saturation deficit, and of precipitation exceeding 0.1 mm h^{-1} (second column of Fig. 11), the $N^{-1/2}$ scaling appears to hold for ensemble sizes of 100 or greater, but the deviations for smaller ensemble sizes are often significant. For the 95th percentile and probability of precipitation exceeding 5.0 mm h^{-1} (third column of Fig. 11), the $N^{-1/2}$ scaling is convincing only for the neighborhood method, which has a larger effective ensemble size than the estimates based on a single grid point.

4 | CONCLUSIONS AND DISCUSSION

Probabilistic weather forecasts are limited by cost to small ensemble sizes. The errors resulting from these small sizes are difficult to assess since we do not know the distribution of a forecast variable that a large ensemble would see, and how many ensemble members would be required to accurately sample it. In this study, we have examined a 1,000-member convection-permitting ensemble forecast to determine what types of distribution arise for different forecast variables, and how well these distributions are represented by smaller ensembles of different sizes. We also consider the performance of forecasts based on neighborhood statistics or area-averaged quantities.

When interpreting the results of this investigation, a number of limitations need to be kept in mind. The forecasts are for a single period of convective weather over a specific geographical region. The results are presented for a single forecast, which is typical of this period, but may not be typical of other regions or weather types. The ensemble has only 1,000 members, which we have found to be sufficient to represent some forecast quantities such as ensemble means, but not others that are more sensitive to extreme events. Furthermore, for practical reasons not all forecast quantities of interest were examined. Some particularly interesting variables to explore would be the probability of precipitation exceeding a threshold, which depends strongly on the number of members with zero precipitation, or near-surface winds and temperatures where interactions with the boundary constrain the forecast distributions. Despite these limitations, we believe that some of our results can be applied more generally, and we focus on these in the following discussion.

The main conclusions of this work are:

1. Three distribution types. All of the histograms produced by the 1,000-member ensemble fell into three categories: quasi-normal, highly skewed, and mixture. These are very general and loosely defined categories, so it is also important to note what was not observed. In particular, the distributions are well-behaved, in the sense that there was no evidence of power-law tails or other extreme forms in Figures 7–9, which might prevent the estimates for a forecast quantity from converging with increasing ensemble size (Fig. 10).

   This conclusion is likely to depend on the space and time-scales of the forecast. In a short-range, limited-area forecast, such as those considered here, the initial and boundary conditions strongly constrain the forecast, and in particular the probability of outlier events. Longer-range forecasts covering a larger spatial domain would allow a wider range of events, and might need larger ensemble sizes to estimate their probabilities. It is possible that as long as predictability is not lost, that is the forecast is constrained by the initial conditions, the forecast distribution will be constrained to fall into the three categories found here, but this can only be determined by further experimentation.
2. Universal asymptotic convergence law. For "well-behaved" forecast quantities, which, as noted above, seems to include all those considered in the present study, we expect the estimates to converge with ensemble size. In the limit of large ensemble size N, the distribution of the estimates for different samples of a given size should approach a Gaussian form, with the width of the confidence interval decreasing as $N^{-1/2}$. This scaling law was indeed found in the

1,000-member ensemble for all three of the example variables examined here, for some forecast quantities. The ensemble mean always followed the expected scaling as did the standard deviation for sufficiently large ensemble size. The scaling law was not unambiguously observed for the 95th percentile or probability of high precipitation rates, except where the neighborhood method was used to increase the effective ensemble size beyond 1,000. The precise ensemble size required to give a desired accuracy (size of confidence interval) will be specific to the forecast being considered, but once the ensemble is large enough to follow the $N^{-1/2}$ scaling law, the available results can be extrapolated to estimate the required ensemble size for any other given level of accuracy.

The results of this study suggest that in deciding what size of ensemble is needed for a particular forecasting problem, it is important to consider whether the ensemble is large enough for the asymptotic convergence behavior to be established. It is significant that for the forecast problems considered here, the ensemble sizes of 40–100 currently used in operational weather forecasting are more than adequate to show convergence of the ensemble mean, and in most cases sufficient for the standard deviation, although clearly inadequate for more extreme events such as the 95th percentile. It is tempting to speculate that, since the operational ensembles are often evaluated in terms of their standard deviation (RMS spread), relative to the mean error, the current ensemble sizes have been chosen as the minimum necessary to give a useful estimate of spread.

3. Neighbourhood and averaging methods can be effective. The results of Section 3.2 show that both the neighborhood and averaging methods can reduce the sampling error associated with small ensemble size. Both methods depend on the small-scale variability of the forecast quantity being uncorrelated in space. For the neighborhood method, the additional points in the neighborhood then provide independent realisations of the variability, leading to a larger effective ensemble size. A quantitative estimate of this increase for ensemble mean precipitation or humidity showed that an increase in effective ensemble size corresponds to a correlation length of about 10 grid points, which is larger than the effective resolution of the model and may be evidence of some degree of convective organization. Forecasts of area-averaged quantities, on the other hand, gain accuracy by averaging out the uncorrelated spatial variability within the averaging region.

Since both of these methods rely on random variability in space, their success for the convection forecasts considered here cannot be generalized to other phenomena such as fog or synoptic weather systems which have smoother spatial structures. Even for 500-hPa temperature in the current forecasts, the two methods brought no benefit. This suggests that convective-scale ensemble forecasting systems may be able to use smaller ensemble sizes than the global systems used for medium-range forecasting. It is also possible that sub seasonal to seasonal forecasts, where the synoptic weather systems can sometimes be regarded as small-scale noise, would again benefit from averaging or neighborhood methods.

The most important conclusion of this work is to recast the question of ensemble size in terms of the asymptotic convergence behavior of forecast quantities. While one can conceive of theoretical distributions that will not show this behavior, we are not aware of any evidence that they arise in practical ensemble prediction problems. Indeed, for such a distribution, one would need to define in what sense it is predictable at all.

More important from a practical point of view, is the question of how to demonstrate and measure convergence without starting with a very large ensemble as a prerequisite. The obvious approach would be to compute confidence intervals (using bootstrapping) for sub-ensembles of different sizes, up to the currently available size. If, as speculated above, current ensembles are near the minimum size required to show scaling for the standard deviation, experiments using a modest increase in size, e.g. two or three times, might be enough to show convergence.

AUTHOR CONTRIBUTIONS

George Craig: Conceptualization; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; validation; writing – original draft; writing – review and editing. **Matjaž Puh:** Formal analysis; investigation; methodology; software; validation; visualization; writing – review and editing. **Christian Keil:** Formal analysis; funding acquisition; investigation; project administration; resources; supervision; validation; writing – review and editing. **Kirsten Tempest:** Conceptualization; formal analysis; methodology; writing – review and editing. **Tobias Necker:** Data curation; methodology; resources; validation; writing – review and editing. **Juan Jose Ruiz:** Data curation; resources; validation; writing – review and editing. **Martin Weissmann:** Data curation; funding acquisition; project administration; resources; writing – review and editing. **Takemasa Miyoshi:** Data curation; methodology; project administration; resources; writing – review and editing.

ORCID

George C. Craig https://orcid.org/0000-0002-7431-8164
Christian Keil https://orcid.org/0000-0003-2736-4309

REFERENCES

Bachmann, K., Keil, C., Craig, G.C., Weissmann, M. and Welzbacher, C.A. (2020) Predictability of deep convection in idealized and operational forecasts: Effects of radar data assimilation, orography, and synoptic weather regime. *Monthly Weather Review*, 148, 63–81. https://doi.org/10.1175/MWR-D-19-0045.1.

Bannister, R.N., Migliorini, S., Rudd, A.C. and Baker, L.H. (2017) Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble. *Geoscientific Model Development Discussions*, 2017, 1–38 URL: https://www.geosci-model-dev-discuss.net/gmd-2017-260/.

Beljaars, A.C.M. and Holtslag, A.A.M. (1991) Flux parameterization over land surfaces for atmospheric models. *Journal of Applied Meteorology*, 30, 327–341. https://doi.org/10.1175/1520-0450(1991)030<0327:FPOLSF>2.0.CO;2.

Bouallègue, Z.B., Theis, S.E. and Gebhardt, C. (2013) Enhancing cosmo-de ensemble forecasts by inexpensive techniques. *Meteorologische Zeitschrift*, 22, 49–59. https://doi.org/10.1127/0941-2948/2013/0374.

Bonamente, M. (2017) *Statistics and Analysis of Scientific Data*. New York: Springer. https://doi.org/10.1007%2F978-14939-6572-4.

Bouttier, F., Vié, B., Nuissier, O. and Raynaud, L. (2012) Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review*, 140, 3706–3721.

Clark, A. J., Gallus Jr, W. A., Xue, M. and F., K. (2010) Growth of spread in convection-allowing and convectionparameterizing ensembles. *Weather and Forecasting*, 25, 594–612. https://doi.org/10.1175/2009WAF2222318.1.

Craig, G.C., Fink, A.H., Hoose, C., Janjić, T., Knippertz, P., Laurian, A., Lerch, S., Mayer, B., Miltenberger, A., Redl, R., Riemer, M., Tempest, K.I. and Wirth, V. (2021) Waves to weather: Exploring the limits of predictability of weather. *Bulletin of the American Meteorological Society*, 102, E2151–E2164. URL: https://journals.ametsoc.org/view/journals/bams/aop/BAMS-D-200035.1/BAMS-D-20-0035.1.xml.

Dekking, F., Kraaikamp, C., Lopuhää, H. and Meester, L. (2005) *A modern introduction to probability and statistics*. London: Springer.

Ebert, E.E. (2009) Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, 24, 1498–1510. https://doi.org/10.1175/2009waf2222251.1.

Evensen, G. and van Leeuwen, P.J. (2000) An ensemble kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128, 1852–1867. https://doi.org/10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2.

Frogner, I.-L., Singleton, A.T., Køltzow, M.Ø. and Andrae, U. (2019) Convection-permitting ensembles: Challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society*, 145, 90–106. https://doi.org/10.1002/qj.3525.

Gebhardt, C., Theis, S.E., Paulat, M. and Ben Bouallègue, Z. (2011) Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, 100, 168–177.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N. and Tennant, W. (2017) The met office convective-scale ensemble, mogreps-uk. *Quarterly Journal of the Royal Meteorological Society*, 143, 2846–2861. https://doi.org/10.1002/qj.3135.

Harnisch, F. and Keil, C. (2015) Initial Conditions for Convective-Scale Ensemble Forecasting Provided by Ensemble Data Assimilation. *Monthly Weather Review*, 143, 1583–1600. https://doi.org/10.1175/MWR-D-14-00209.1.

Hirt, M., Rasp, S., Blahak, U. and Craig, G.C. (2019) Stochastic Parameterization of Processes Leading to Convective Initiation in Kilometer-Scale Models. *Monthly Weather Review*, 147, 3917–3934.

Hohenegger, C., Lüthi, D. and Schär, C. (2006) Predictability mysteries in cloud-resolving models. *Monthly Weather Review*, 134, 2095–2107. https://doi.org/10.1175/MWR3176.1.

Hohenegger, C. and Schär, C. (2007) Predictability and error growth dynamics in cloud-resolving models. *Journal of the Atmospheric Sciences*, 64, 4467–4478 URL: https://journals.ametsoc.org/view/journals/atsc/64/12/2007jas2143.1.xml.

Jacques, D. and Zawadzki, I. (2015) The impacts of representing the correlation of errors in radar data assimilation. part ii: Model output as background estimates. *Monthly Weather Review*, 143, 2637–2656. https://doi.org/10.1175/MWRD-14-00243.1.

Jankov, I., Berner, J., Beck, J., Jiang, H., Olson, J.B., Grell, G., Smirnova, T.G., Benjamin, S.G. and Brown, J.M. (2017) A performance comparison between multiphysics and stochastic approaches within a North American RAP ensemble. *Monthly Weather Review*, 145, 1161–1179.

Kawabata, T. and Ueno, G. (2020) Non-gaussian probability densities of convection initiation and development investigated using a particle filter with a storm-scale numerical weather prediction model. *Monthly Weather Review*, 148, 3–20. https://doi.org/10.1175/MWR-D-18-0367.1?mobileUi=0.

Keil, C., Baur, F., Bachmann, K., Rasp, S., Schneider, L. and Barthlott, C. (2019) Relative contribution of soil moisture, boundarylayer and microphysical perturbations on convective predictability in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, 145, 3102–3115. https://doi.org/10.1002/qj.3607.

Keil, C., Chabert, L., Nuissier, O. and Raynaud, L. (2020) Dependence of predictability of precipitation in the northwestern mediterranean coastal region on the strength of synoptic control. *Atmospheric Chemistry and Physics*, 20, 15851–15865. https://doi.org/10.5194/acp-20-15851-2020.

Keil, C., Heinlein, F. and Craig, G.C. (2014) The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of Royal Meteorological Society*, 140, 480–490.

Kondo, K. and Miyoshi, T. (2019) Non-Gaussian statistics in global atmospheric dynamics: a study with a 10 240-member ensemble Kalman filter using an intermediate atmospheric general circulation model. *Nonlinear Processes in Geophysics*, 26, 211–225 URL: https://www.nonlin-processes-geophys.net/26/211/2019/.

Kühnlein, C., Keil, C., Craig, G.C. and Gebhardt, C. (2014) The impact of downscaled initial condition perturbations on convective-scale ensemble forecasts of precipitation. *Quarterly Journal of Royal Meteorological Society*, 140, 1552–1562.

Legrand, R., Michel, Y. and Montmerle, T. (2016) Diagnosing non-gaussianity of forecast and analysis errors in a convectivescale model. *Nonlinear Processes in Geophysics*, 23, 1–12 URL: https://npg.copernicus.org/articles/23/1/2016/.

Leoncini, G., Plant, R.S., Gray, S.L. and Clark, P.A. (2010) Perturbation growth at the convective scale for csip iop18. *Quarterly Journal of the Royal Meteorological Society*, 136, 653–670. https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.587.

Leutbecher, M. (2019) Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, 145, 107–128. https://doi.org/10.1002/qj.3387.

Lien, G.-Y., Miyoshi, T., Nishizawa, S., Yoshida, R., Yashiro, H., Adachi, S.A., Yamaura, T. and Tomita, H. (2017) The near-realtime scale-letkf system: A case of the september 2015 kanto-tohoku heavy rainfall. *SOLA*, 13, 1–6.

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21, 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.

Marsigli, C., Montani, A. and Paccagnella, T. (2014) Perturbation of initial and boundary conditions for a limited-area ensemble: multi-model versus single-model approach. *Quarterly Journal of the Royal Meteorological Society*, 140, 197–208. https://doi.org/10.1002/qj.2128.

Miyoshi, T., Kondo, K. and Imamura, T. (2014) The 10,240-member ensemble kalman filtering with an intermediate agcm. *Geophysical Research Letters*, 41, 5264–5271. https://doi.org/10.1002/2014GL060863.

Montani, A., Cesari, D., Marsigli, C. and Paccagnella, T. (2011) Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A: Dynamic Meteorology and Oceanography*, 63, 605–624. https://doi.org/10.1111/j.1600-0870.2010.00499.x.

Nakanishi, M. and Niino, H. (2004) An improved mellor–yamada level-3 model with condensation physics: Its design and verification. *Boundary-Layer Meteorology*, 112, 1–31. https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

Necker, T., Geiss, S., Weissmann, M., Ruiz, J., Miyoshi, T. and Lien, G.-Y. (2020a) A convective-scale 1,000-member ensemble simulation and potential applications. *Quarterly Journal of the Royal Meteorological Society*, 146, 1423–1442. https://doi.org/10.1002/qj.3744.

Necker, T., Weissmann, M., Ruckstuhl, Y., Anderson, J. and Miyoshi, T. (2020b) Sampling Error Correction Evaluated Using a Convective-Scale 1000-Member Ensemble. *Monthly Weather Review*, 148, 1229–1249. https://doi.org/10.1175/MWR-D-19-0154.1.

Necker, T., Weissmann, M. and Sommer, M. (2018) The importance of appropriate verification metrics for the assessment of observation impact in a convection-permitting modelling system. *Quarterly Journal of Royal Meteorological Society*, 144, 1667–1680.

Nishizawa, S. and Kitamura, Y. (2018) A surface flux scheme based on the monin-obukhov similarity for finite volume models. *Journal of Advances in Modeling Earth Systems*, 10, 3159–3175. https://doi.org/10.1029/2018MS001534.

Nishizawa, S., Yashiro, H., Sato, Y., Miyamoto, Y. and Tomita, H. (2015) Influence of grid aspect ratio on planetary boundary layer turbulence in large-eddy simulations. *Geoscientific Model Development*, 8, 3393–3419. https://gmd.copernicus.org/articles/8/3393/2015/.

Palmer, T.N., Gelaro, R., Barkmeijer, J. and Buizza, R. (1998) Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences*, 55, 633–653. https://doi.org/10.1175/1520-0469(1998)055<0633:svmaao>2.0.co;2.

Piper, D., Kunz, M., Ehmele, F., Mohr, S., Mühr, B., Kron, A. and Daniell, J. (2016) Exceptional sequence of severe thunderstorms and related flash floods in May and June 2016 in Germany – part 1: Meteorological background. *Natural Hazards and Earth System Sciences*, 16, 2835–2850 URL: https://www.nat-hazards-earth-syst-sci.net/16/2835/2016/.

Rasp, S., Selz, T. and Craig, G.C. (2018) Variability and clustering of midlatitude summertime convection: testing the Craig and Cohen theory in a convection-permitting ensemble with stochastic boundary layer perturbations. *Journal of Atmosphere Science*, 75, 691–706.

Raynaud, L. and Bouttier, F. (2017) The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 143, 3037–3047. https://doi.org/10.1002/qj.3159.

Ruiz, J., Lien, G.-Y., Kondo, K., Otsuka, S. and Miyoshi, T. (2021) Reduced non-gaussianity by 30-second rapid update in convective-scale numerical weather prediction. *Nonlinear Processes in Geophysics*, 28, 615–626. https://doi.org/10.5194/npg-28-615-2021.

Sakradzija, M., Senf, F., Scheck, L., Ahlgrimm, M. and Klocke, D. (2020) Local impact of stochastic shallow convection on clouds and precipitation in the tropical atlantic. *Monthly Weather Review*, 148, 5041–5062. https://doi.org/10.1175/mwrd-20-0107.1.

Sato, Y., Nishizawa, S., Yashiro, H., Miyamoto, Y., Kajikawa, Y. and Tomita, H. (2015) Impacts of cloud microphysics on trade wind cumulus: which cloud microphysics processes contribute to the diversity in a large eddy simulation? *Progress in Earth and Planetary Science*, 2, 23. https://doi.org/10.1186/s40645-015-0053-6.

Scheck, L., Weissmann, M. and Bach, L. (2020) Assimilating visible satellite images for convective-scale numerical weather prediction: A case-study. *Quarterly Journal of the Royal Meteorological Society*, 146, 3165–3186. https://doi.org/10.1002/qj.3840.

Schwartz, C.S., Romine, G.S., Fossell, K.R., Sobash, R.A. and Weisman, M.L. (2017) Toward 1-km ensemble forecasts over large domains. *Monthly Weather Review*, 145, 2943–2969. https://doi.org/10.1175/mwr-d-16-0410.1.

Sekiguchi, M. and Nakajima, T. (2008) A k-distribution-based radiation code and its computational optimization for an atmospheric general circulation model. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 109, 2779–2793 URL: http://www.sciencedirect.com/science/article/pii/S0022407308001635.

Selz, T. (2019) Estimating the intrinsic limit of predictability using a stochastic convection scheme. *Journal of the Atmospheric Sciences*, 76, 757–765. https://doi.org/10.1175/jas-d-17-0373.1.

Stensrud, D.J., Xue, M., Wicker, L.J., Kelleher, K.E., Foster, M.P., Schaefer, J.T., Schneider, R.S., Benjamin, S.G., Weygandt, S.S., Ferree, J.T. and Tuell, J.P. (2009) Convective-scale warn-on-forecast system. *Bulletin of the American Meteorological Society*, 90, 1487–1500. https://doi.org/10.1175/2009BAMS2795.1.

Sun, J., Xue, M., Wilson, J.W., Zawadzki, I., Ballard, S.P., Onvlee-Hooimeyer, J., Joe, P., Barker, D.M., Li, P.-W., Golding, B., Xu, M. and Pinto, J. (2014) Use of nwp for nowcasting convective precipitation: Recent progress and challenges. *Bulletin of the American Meteorological Society*, 95, 409–426. https://doi.org/10.1175/BAMS-D-11-00263.1.

Tomita, H. (2008) New microphysical schemes with five and six categories by diagnostic generation of cloud ice. *Journal of the Meteorological Society of Japan. Ser. II*, 86A, 121–142.

Toth, Z. and Buizza, R. (2019) Chapter 2 - weather forecasting: What sets the forecast skill horizon? In *Sub-Seasonal to Seasonal Prediction* (eds. A. W. Robertson and F. Vitart), 17–45. Amsterdam: Elsevier. URL: https://www.sciencedirect.com/science/article/pii/B9780128117149000024.

Toth, Z. and Kalnay, E. (1997) Ensemble forecasting at NCEP and the breeding method. *Monthly Weather Review*, 125, 3297–3319. https://doi.org/10.1175/1520-0493(1997)125<3297:efanat>2.0.co;2.

Walker, A.M. (1968) A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society: Series B (Methodological)*, 30, 570–575. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1968.tb00757.x.

Zhang, F. (2005) Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts. *Monthly Weather Review*, 133, 2876–2893. https://doi.org/10.1175/MWR3009.1.

How to cite this article: Craig, G.C., Puh, M., Keil, C., Tempest, K., Necker, T., Ruiz, J. *et al*. (2022) Distributions and convergence of forecast variables in a 1,000-member convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*, 148(746), 2325–2343. Available from: https://doi.org/10.1002/qj.4305