Received: 18 August 2022 | Revised: 18 November 2022 | Accepted: 21 November 2022 | Published on: 22 March 2023

DOI: 10.1002/qj.4410

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble

Kirsten I. Tempest[1] | **George C. Craig**[1] | **Jonas R. Brehmer**[2]

[1]Meteorological Institute, Ludwig-Maximilian-University Munich, Munich, Germany

[2]Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

Correspondence
K. I. Tempest, Meteorological Institute, Ludwig-Maximilian-University Munich, Munich, Germany.
Email: K.Tempest@physik.uni-muenchen.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Number: SFB / TRR 165; Klaus Tschira Stiftung

Abstract

Many operational weather services use ensembles of forecasts to generate probabilistic predictions. Computational costs generally limit the size of the ensemble to fewer than 100 members, although the large number of degrees of freedom in the forecast model would suggest that a vastly larger ensemble would be required to represent the forecast probability distribution accurately. In this study, we use a computationally efficient idealised model that replicates key properties of the dynamics and statistics of cumulus convection to identify how the sampling uncertainty of statistical quantities converges with ensemble size. Convergence is quantified by computing the width of the 95% confidence interval of the sampling distribution of random variables, using bootstrapping on the ensemble distributions at individual time and grid points. Using ensemble sizes of up to 100,000 members, it was found that for all computed distribution properties, including mean, variance, skew, kurtosis, and several quantiles, the sampling uncertainty scaled as $n^{-1/2}$ for sufficiently large ensemble size n. This behaviour is expected from the Central Limit Theorem, which further predicts that the magnitude of the uncertainty depends on the distribution shape, with a large uncertainty for statistics that depend on rare events. This prediction was also confirmed, with the additional observation that such statistics also required larger ensemble sizes before entering the asymptotic regime. By considering two methods for evaluating asymptotic behaviour in small ensembles, we show that the large-n theory can be applied usefully for some forecast quantities even for the ensemble sizes in operational use today.

KEYWORDS

asymptotic convergence, distributions, ensembles, idealised model, sampling uncertainty, weather prediction

1 | INTRODUCTION

Probabilistic forecasting is currently used by many operational forecasting facilities. In comparison with deterministic forecasting, it provides important benefits. Probabilistic forecasting allows for flow-dependent uncertainty to be quantified and evolved in time, which in turn allows for a probability to be attached to the meteorological prediction (Leutbecher and Palmer, 2008). As a result, the economic value of a probabilistic forecast is generally higher than that of a deterministic forecast (Zhu *et al.*, 2002).

The quality of probabilistic forecasts has improved steadily in recent years (Bauer *et al.*, 2015), through a combination of improvements to the observing system and data assimilation (DA) methods as well as the formulation of numerical models, including the representation of model error through stochastic parameterisations (e.g. Bouttier *et al.*, 2012; Jankov *et al.*, 2017; Rasp *et al.*, 2017; Hirt *et al.*, 2019; Sakradzija *et al.*, 2020). On the other hand, it is well known that initial condition and model errors, combined with the chaotic nature of the atmosphere, lead to an intrinsic limit on the predictability of the atmosphere (Lorenz, 1969). However, the error-growth experiments of Selz and Craig (2021) suggest that an additional 4–5 days of predictability can still be gained by further improvements to the forecasting system.

Despite huge increases in computing power, one aspect of the forecasting system that has not changed much is ensemble size. Operational ensembles typically have sizes of 20–50 members (Buizza *et al.*, 2000; Reinert *et al.*, 2020; Met Office, 2022), and apparently the improvements in forecast skill that might be obtained from a larger ensemble do not justify the costs of more members. Buizza *et al.* (1998) and Raynaud and Bouttier (2017) compared the benefits of increased ensemble size with those of higher resolution for global and regional forecasting systems, respectively. Both studies found that either ensemble size or resolution increases could be more beneficial, depending on factors such as forecast lead time and the quantity being predicted. Other improvements such as a better quantification of the initial and model uncertainties may also improve the forecast. Leutbecher (2019) provides a theoretical framework for the modest increases in forecast skill with increasing ensemble size. A number of different skill scores were evaluated, and results from European Centre for Medium-Range Weather Forecasts (ECMWF) ensembles with up to 200 members were compared with theoretical expectations for ensembles of different sizes under the assumptions that the ensemble is reliable and the members are exchangeable. Under these assumptions, for example, the continuous ranked probability score (CRPS) of an ensemble of size n is equal to the score for an infinite ensemble multiplied by a factor $(1 + 1/n)$. This shows that improvements in CRPS will be small once the ensemble size has reached a few tens of members, and useful estimates can often be obtained with even smaller ensembles. Similar results were found for other scores, with the notable exception of the quantile score (QS) for the more extreme quantiles close to 0 or 1, where convergence required much larger ensemble sizes. As shown by Richardson (2001), such forecasts are particularly important to users with low cost/loss ratios.

The sensitivity for extreme quantiles is perhaps unsurprising, since the frequency distribution of a forecast quantity (hereafter referred to simply as the distribution) from an ensemble of up to 50 members is unlikely to be accurate for rare events that are sampled infrequently. In research environments, larger ensemble sizes have been considered. For example, Lin *et al.* (2020) evaluated a measure of hurricane strength, nondimensional damage, that depends nonlinearly on wind speed and is sensitive to extremes. They found that a 100-member ensemble was not large enough to resolve the relevant part of the wind-speed distribution, whereas an ensemble size of 1,000 gave much improved results. Likewise, Jacques and Zawadzki (2015) chose to use a 1,000-member ensemble to describe the background covariance structure required for DA, since multivariate combinations of values may be sampled infrequently even when the individual values are not rare. This effect is magnified further by the fact that multivariate distributions may have extremely non-Gaussian properties, even when derived from quasi-Gaussian marginals (Poterjoy, 2022). Such behaviour is particularly problematic in DA, where distributions are often assumed to be multivariate Gaussian in form. A quantitative evaluation of the importance of ensemble size in DA was provided by Kondo and Miyoshi (2019), who used the 10,240 member global ensemble of Miyoshi *et al.* (2014) and measured the degree of non-Gaussianity for different ensemble sizes. It was found that approximately 1,000 members were generally required to represent characteristics of non-Gaussian distributions such as skewness and kurtosis.

The preceding studies show that the skill of ensemble predictions depends on the resolution of the distribution produced by the forecast ensemble, where the ensemble size is of direct importance, but also on its reliability, how well the ensemble reproduces observations, and the user requirements that determine which properties of the distribution are of interest. In this article, we will investigate the dependence of sampling uncertainty on ensemble size, and leave questions of reliability and score against observations for future work. This will allow us to consider the effects of the different distribution shapes that arise for

different forecast variables and lead times, and provides a basis for future work to quantify the contributions of different sources of uncertainty in the ensemble. The present study builds on the results from a 1,000-member convective-scale ensemble (Craig *et al.*, 2022), but uses an idealised model where it is possible to use very large ensemble sizes to examine the convergence of any variable of interest.

A first idea of how forecast distribution shapes vary as a function of lead time is given by the conceptual model proposed by Craig *et al.* (2022) (see their figure 1). Most DA systems assume that the distributions are close to Gaussian during the assimilation cycle. In the free forecast, the distribution will broaden as uncertainty increases and will start to deviate from Gaussianity as nonlinear processes become important. The shape of the distribution may develop long tails (extreme events) and multiple peaks (weather regimes). At long lead times, it will converge to a smoother climatological distribution. Looking at forecasts of winds, temperatures, cloud, and precipitation from a 1,000-member convective-scale ensemble, Craig *et al.* (2022) found evidence for this progression, but, perhaps more importantly, they found that all the distributions they observed for different variables and lead times could be categorised into three basic shapes: quasi-Gaussian, highly skewed, and multimodal, with different convergence properties. These results are limited, however, by the data set available, which consists of a small number of forecast days for a particular season and geographical region.

The most interesting result of Craig *et al.* (2022) was that, for all distribution shapes and most forecast variables, the width of confidence intervals of ensemble estimates decreased with increasing ensemble size n following a universal scaling law $n^{-1/2}$. The forecast variables included the mean, standard deviation, and 95th percentile of temperature and humidity, as well as the probability of precipitation exceeding certain thresholds. The $n^{-1/2}$ scaling was found for sufficiently large n for all quantities, except some 95th percentiles and probability of precipitation exceeding large thresholds. It is possible that the scaling would eventually be observed for ensemble sizes larger than 1,000 members. This convergence behaviour is expected from the Central Limit Theorem (CLT), whereby, for a large number n of independent and identically distributed (iid) random variables, the sampling distribution of the summation of the random variables will be normally distributed without dependence on the initial distribution shape (Dekking *et al.*, 2005). Furthermore, the standard error of this sampling distribution will be proportional to $n^{-1/2}$ in the limit of large n. Convergence in the uncertainty of the ensemble mean with ensemble size according to this theory was illustrated by Leutbecher (2019) for a Gaussian distribution and for an example of 500-hPa geopotential over the Northern Hemisphere.

While the $n^{-1/2}$ scaling of the uncertainty with ensemble size is independent of the underlying distribution, the absolute magnitude is not. For a given quantile level, the standard deviation of the ensemble sample estimate of that quantile is given by

$$\sigma_{n_p} = \frac{1}{\sqrt{n}} \sqrt{\frac{p(1-p)}{f^2(q_p)}}, \tag{1}$$

where σ_{n_p} is the standard deviation of the quantile sampling distribution, n is the number of ensemble members, and f is the probability density at q_p, the true theoretical quantile corresponding to p, where $p \in (0, 1)$ is the quantile level (Stuart and Ord, 2000; Gneiting, 2014). The first term on the right-hand side of Equation 1 shows the expected scaling with ensemble size, while the second term shows that the uncertainty is inversely proportional to the frequency of occurrence of the quantile, that is, predictions of rare events are less confident. For sufficiently large n, Equation 1 provides an estimate of how many ensemble members would be required in order to have a specific level of sampling uncertainty for a particular quantile level of a meteorological variable, and how this changes depending on quantile level. This is illustrated in Figure 1, which shows the ensemble size required to reach a given level of sampling uncertainty for different quantile levels for a Gaussian-distributed variable, computed from Equation 1. The figure shows that, as one requires increased certainty in the estimate of the quantile level p, more members are required. Furthermore, as the quantile level gets more extreme (further away from the median), the uncertainty increases for any given number of ensemble members, varying inversely with the underlying Gaussian distribution shown in Figure 1a. It is unknown whether this equation will be useful for meteorological data, however, due to not knowing the exact underlying probability density function (PDF) and not necessarily having sufficient ensemble members for the asymptotic results to be accurate.

In cases where asymptotic scaling is actually observed, it should be possible to approximate the number of ensemble members required to reach a given level of sampling uncertainty for a statistical quantity of a forecast variable. For example, if a forecaster wants to approximate the number of members required to reach a certain accuracy in the spread of a measurement of temperature over Munich, asymptotic scaling could be used. This would work by quantifying how the data the forecaster has with the current sized ensemble scales with $n^{-1/2}$, and then extrapolating this until it reaches the level of sampling uncertainty that is wanted.

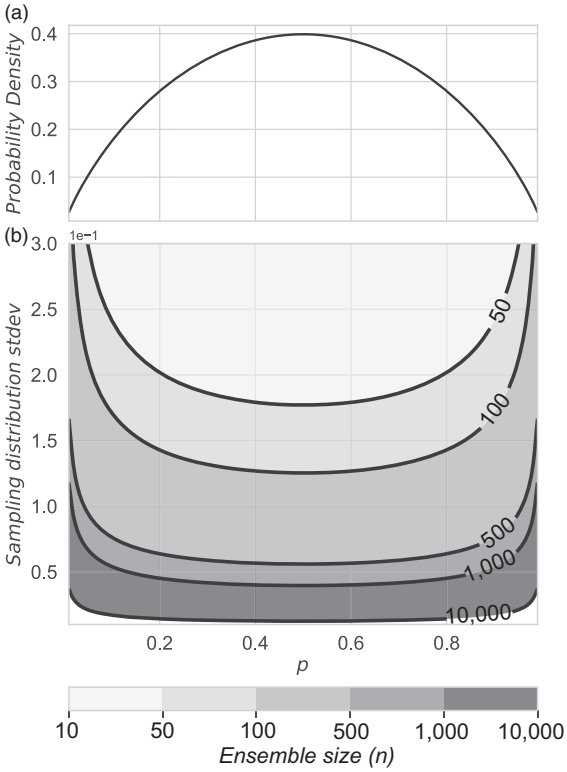

FIGURE 1 (a) Probability density of a Gaussian as a function of quantile p and (b) showing the corresponding ensemble size (contours) required to reach a given level of sampling uncertainty (y-axis) for different quantile levels (x-axis)

In this article, we assess the relevance of the asymptotic theory to ensemble weather prediction by considering a computationally efficient idealised ensemble forecasting system. We will investigate the ensemble sizes required to obtain the asymptotic scaling law for different quantities and their dependence on the underlying distribution. Finally, we will consider how to obtain information about convergence from ensembles of limited size.

In Section 2, the model and methods are presented. An idealised model is selected and the setup of the idealised prediction system is described, along with the methods that are carried out on the subsequent ensemble data. The first part of Section 3 evaluates whether the distributions from the idealised prediction system are of a form similar to those from Craig *et al.* (2022). The results of exploring the convergence behaviour are reported in Section 3.2. In Section 3.3, two methods of estimating the sampling uncertainty at larger ensemble sizes using only a smaller ensemble are discussed. The main results are summarised in the conclusions in Section 4.

2 | MODEL AND METHODS

For this study, a model is required which represents the basic processes of convection in the midlatitude atmosphere. This encompasses having spatial and time scales representative of convective processes and being capable of modelling nonlinear processes. It must, in addition, be computationally inexpensive, so that ensemble sizes of order $\mathcal{O}(10^5)$ can be examined efficiently. We employ a one-dimensional idealised model for cumulus convection (Würsch and Craig, 2014, hereafter referred to as WC14), which was developed for convective-scale DA. This model features a simple representation of convective updrafts and downdrafts, but with enough complexity to mimic the nonlinear dynamics of the convective life cycle and the spatially intermittent and non-Gaussian statistics of a convecting atmosphere. In Section 2.1 the model of WC14 is described, and in Section 2.2 we assess whether this model achieves the requirements stated above. The idealised prediction system built on the basis of the idealised model is presented in Section 2.3, before the methods used are outlined in Section 2.4.

2.1 | Model

The one-dimensional idealised model (WC14) uses a modified version of the shallow-water equations for a single fluid layer. Conditional instability that leads to convection is modelled by a modification of the buoyancy term when the fluid level is lifted sufficiently, and a rain equation is introduced to allow for the creation of negatively buoyant downdrafts. The model state is specified by three variables: wind u, height h, and rain r, illustrated in Figure 2. These are described by the following equations:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial(\phi + c^2 r)}{\partial x} = K_u \frac{\partial^2 u}{\partial x^2} + F, \tag{2}$$

$$Z = h + H, \tag{3}$$

$$\phi = \begin{Bmatrix} \phi_c + gH, & Z > H_c \\ g(H + h), & \text{otherwise} \end{Bmatrix}, \tag{4}$$

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = K_h \frac{\partial^2 h}{\partial x^2}, \tag{5}$$

$$\frac{\partial r}{\partial t} + u\frac{\partial r}{\partial x} = K_r \frac{\partial^2 r}{\partial x^2} - \alpha r - \begin{cases} \beta \frac{\partial u}{\partial x}, & Z > H_r \text{ and } \frac{\partial u}{\partial x} < 0 \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

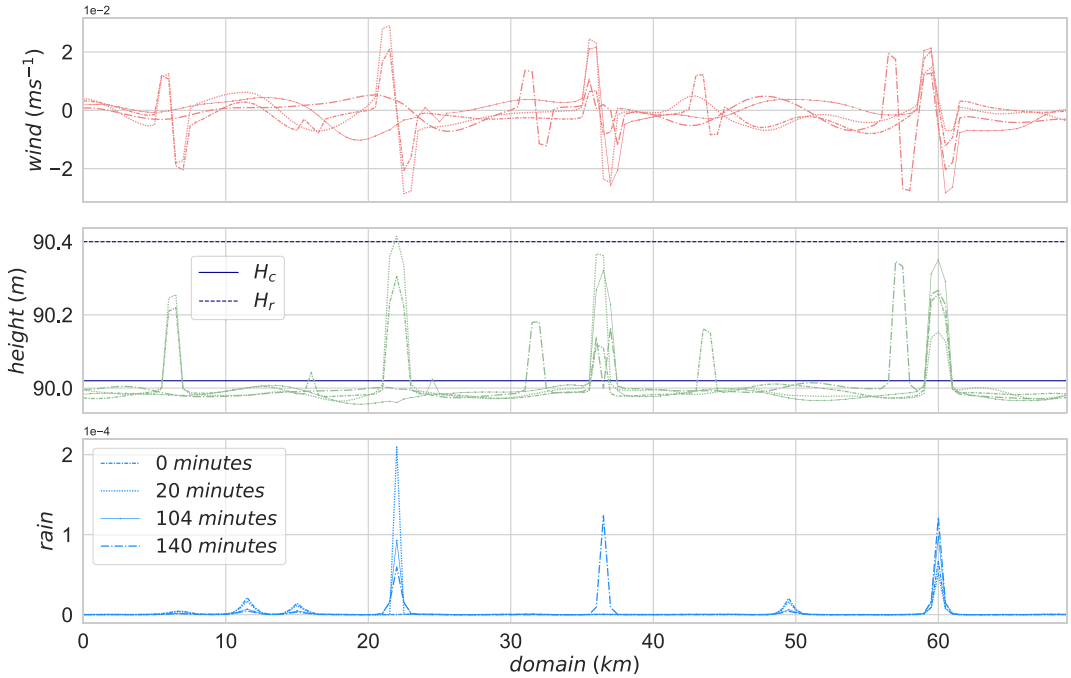

FIGURE 2 Snapshots of the domain at four time points for the three model variables. Thresholds (described in text) are shown in the height field [Colour figure can be viewed at wileyonlinelibrary.com]

where H is the height of the topography, h is the fluid depth (referred to as "height"), and $Z = H + h$ the absolute fluid layer height. Note that in the present study we do not include orography, so that $H = 0$ and therefore $Z = h$. From selecting the initial fluid level height, h_0, to be 90 m, the gravity-wave speed is 30 m $\cdot$ s^{-1}, as in WC14. The diffusion constants used are $K_u = 2 \times 10^3$ m$^2 \cdot$ s^{-1}, $K_h = 6 \times 10^3$ m$^2 \cdot$ s^{-1}, and $K_r = 10$ m$^2 \cdot$ s^{-1}.

If h is greater than a first threshold ($h > H_c = 90.02$ m), then the buoyancy at that grid point is increased by setting the geopotential, ϕ, to a relatively low constant, ϕ_c, which is chosen to be 899.77 m$^2 \cdot$ s^{-2}. This encourages more fluid into this region, thereby increasing h further. This process is analogous to the buoyant updraft phase of a cloud, whereby the level of free convection has been passed by a saturated fluid parcel. Therefore, when h crosses the threshold H_c, that grid point is said to contain a cloud.

If h crosses a second threshold ($h > H_r = 90.4$ m) and wind is converging on this grid point, then rain (scaled by β, which is set to 0.1) is produced. Where rain exists, it adds a negative term to the geopotential, reducing buoyancy and tending to create downward motion, leading to the collapse of the cloud. Rain is removed from the domain by a linear relaxation of rate α, with value $1.4 \times 10^{-4} s^{-1}$. This allows for rain to remain at a grid point even if there is no longer a cloud, thereby disincentivising another cloud to form immediately afterwards at the same location. An example of the growth and decay of a short-lived cloud occurs at $x = 22$ km in Figure 2. The height crosses the rain threshold at $t = 20$ min, the negative buoyancy due to the rain changes the convergent wind to divergent, and the height perturbation has disappeared by $t = 104$ min, while the rain amount decays more gradually.

Throughout the simulation, gravity waves perturb the height field, initiating and inhibiting convection. In addition, to model the contribution of boundary-layer turbulence to convective initiation, convergent and divergent perturbations F are added to the wind field at every time step. These are of the form of a normalised first-order derivative of a Gaussian function. This odd function is multiplied by an amplitude, $\bar{u}$, which has value $8.95 \times 10^{-3} ms^{-1}$. Convergent perturbations encourage h to reach the first threshold in height (H_c), initiating the updraft phase of a cloud.

The numerical implementation of the model is based on WC14, with a second-order centred finite-difference approximation on a staggered grid alongside a RAW filter for time-smoothing (Williams, 2009; 2011). The time step is modified here to 4 s and the RAW filter parameter to 0.7, for numerical stability. The integrated height field over the domain does not change in time, signifying that the model is mass-conserving under this numerical approximation.

2.2 | Properties of the model solutions

The example in Figure 2 shows that the evolution of the simulated cloud life cycle occurs on realistic time scales.

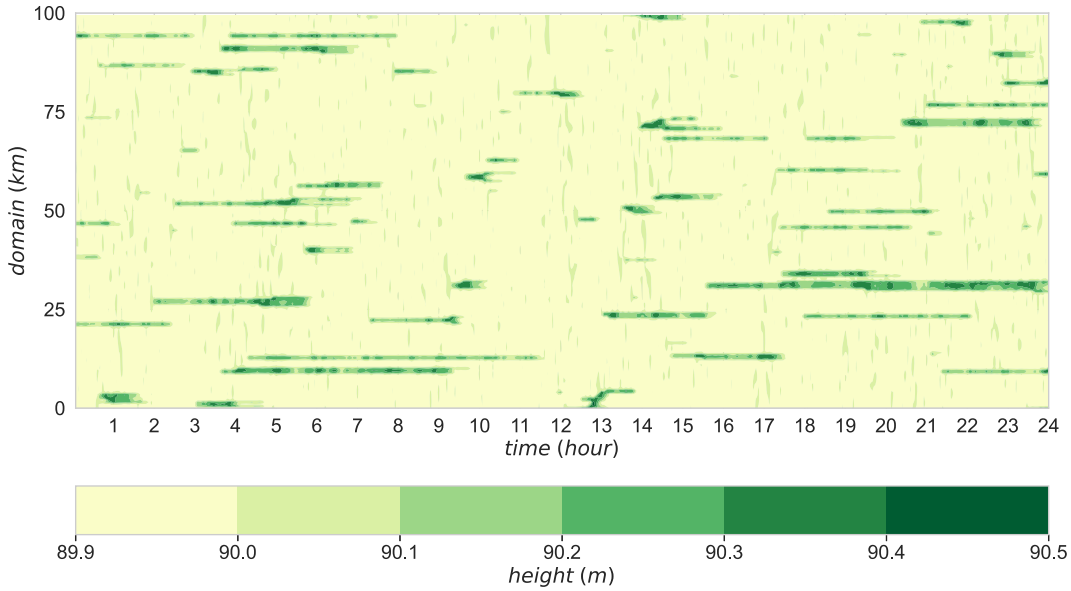

FIGURE 3 Hovmöller diagram showing the evolution of the height field across a section of the domain over the 24-hr period of the free run [Colour figure can be viewed at wileyonlinelibrary.com]

For the updraft phase of a cloud, the time between a cloud's initiation ($h > H_c$) and rain formation ($h > H_r$) is approximately 15 min. For the downdraft phase, the half-life of rain is approximately 1 hr and the overall lifetime of a cloud ($h > H_c$) with one updraft and one downdraft is between 1 and 2 hr. Multiple phases of a cloud can exist, as shown in Figure 3, which displays the evolution of the height field in time using a Hovmöller diagram. Longer lasting clouds exist, featuring several up- and downdraft phases in their evolution (marked by multiple regions with darker shades of green). Splitting of convective updrafts and initiation of new clouds in the vicinity of existing clouds can also be observed.

With a total domain size of 500 km and a horizontal resolution of 500 m, there is a cloud coverage of approximately 5% of the domain at any instant. The widths of clouds are logarithmically distributed, with a mean of 1.2 km and a maximum width of 7.5 km, which is in agreement with WC14. This corresponds to an average of 20.8 clouds in the domain at any given time. The statistics of cloud size and number are stationary in time, and the spatial locations of the clouds are close to random, with the distance between clouds following an exponential distribution (not shown). This agrees with the theory detailed in Craig and Cohen (2006) and the numerical experiments of Cohen and Craig (2006). Overall, the temporal and spatial distributions produced by the idealised model are reasonable for a convecting atmosphere. This, along with the computationally inexpensive nature of the modified shallow-water equations, makes it a suitable model for our experiments.

2.3 | Idealised prediction system

A Numerical Weather Prediction (NWP) system is created, based upon the idealised model. A truth run is initialised, along with a 500 member ensemble for DA, which will be used to initialise a larger forecast ensemble. The truth run and ensemble members are initialised with a homogeneous state of no background wind, no rain, and a constant initial height (h_0) of 90 m, and all simulations are run for 1,000 time steps with independent realisations of the stochastic forcing term to spin up the model fields.

After initialisation, the ensemble Kalman filter (EnKF) DA (Evensen, 1994) is cycled 50 times. Observations were assimilated every 5 min at every grid point for each model variable. The observations were obtained by adding a Gaussian (log-normal) noise to the wind and height (rain) fields. This noise has an error of approximately 10% of the maximum deviation from that variable's mean value. A forecast-error covariance localisation (Gaspari and Cohn, 1999) is further implemented, with a localisation radius of 2 km. For more details on the DA used in this system, see Ruckstuhl and Janjić (2018) and Ruckstuhl *et al.* (2021). After 50 cycles, the root mean square error (RMSE) had converged to an approximately constant value. The DA ensemble size of 500 members was chosen based on the results of Ruckstuhl and Janjić (2018) comparing RMSE as a function of ensemble size.

For the free forecast, the ensemble size was expanded to 100,000 members by copying the initial conditions of the DA 200 times each, as even with an idealised setup it

was prohibitive to run the DA with all 100,000 members. This procedure is sufficient, since the stochastic forcing causes members that start with identical initial conditions to decorrelate rapidly. This was verified by computing the Pearson correlation coefficient of the height field over the domain between ensemble members that started with different initial conditions, compared with those that started with identical initial conditions. The forecast ensemble, as well as the truth run, was run for 24 hr, and data were saved every four model minutes. The ability to run such a large ensemble was the primary motivation for using an idealised model.

The NWP system described here models different sources of forecast error. The EnKF provided initial conditions with an approximately Gaussian error. Along with this, the stochastic perturbations to the wind field provide model error. On the other hand, due to the cyclic domain, there are no boundary condition errors.

2.4 | Statistical analysis

The analysis of the ensemble forecasts will focus on two types of statistics. The shape of the distributions of model variables is of particular interest, along with their divergence from being Gaussian-distributed. Furthermore, the nature of the decrease of sampling uncertainty as an ensemble becomes larger is of importance, for which statistical inference will be employed.

2.4.1 | Non-Gaussian statistics

To test how close the forecast distributions are to being Gaussian-distributed, we employ the same measures as used by Kondo and Miyoshi (2019). These are sample skewness, sample excess kurtosis, and Kullback–Leibler divergence (KL divergence). Skewness, the third moment of the distribution, measures the symmetry of the data. Kurtosis, the fourth moment of the distribution, measures the density at the tails of the data. For a Gaussian distribution, skewness and excess kurtosis are zero. The KL divergence is a direct measure of the difference between two distributions. In this study, the KL divergence is used to measure the distance of a histogram of a distribution from the ensemble from that of a reference Gaussian PDF. As such, the lower the score, the closer the distribution from the ensemble is to being Gaussian-distributed, and a subjective threshold is chosen to determine whether that distribution can then be considered Gaussian. Scores above 0.3 are considered here to be non-Gaussian, which is slightly higher than the threshold used by Kondo and Miyoshi (2019).

2.4.2 | Statistical inference

Each finite-sized data set $(x_1, x_2, \ldots, x_n)$ of length n created by an ensemble with n members is just one realisation of the random variables $(X_1, X_2, \ldots, X_n)$ from a distribution F, and, as such, each of the sample statistics (e.g. sample mean $\overline{x}_n = (x_1 + x_2 + \cdots + x_n)/n$) is just one possible realisation of a random variable (e.g. $\overline{X}_n = (X_1 + X_2 + \cdots + X_n)/n$) (Dekking *et al.*, 2005). For inference of a population characteristic of F that the sample statistic is estimating (in this case the sample mean is estimating the expectation μ), the distribution function of the random variable (in this example $\overline{X}_n$) will determine the associated uncertainty of the estimation.

If this underlying distribution F is unknown, nonparametric bootstrapping (Davison and Hinkley, 1997) is a powerful tool used to infer information about its characteristics. This has been used previously in meteorological applications (e.g. Feng *et al.*, 2011). Bootstrapping assumes that the estimate $\hat{F}$ is an accurate realisation of F. Nonparametric bootstrapping is resampling with replacement from a data set where all data points have equal probabilities $1/n$, to create a "bootstrapped" random sample $(X_1^*, X_2^*, \ldots, X_n^*)$, of the same length as the original sample. From each bootstrapped random sample, the desired sample statistic can be calculated (in this case the bootstrapped sample mean $\overline{x}_n^*$). The distribution of this statistic (the random variable of the bootstrapped mean $\overline{X}_n^*$) can then be used to construct confidence intervals and make inferences for the chosen characteristic of F. Using this probability distribution as an approximation for that of the distribution of a random variable, in this case $\overline{X}_n$, is known as the bootstrap principle (Dekking *et al.*, 2005).

For the analysis of uncertainty in this article, bootstrapping will be performed on the distributions obtained from the forecast ensemble described above. The 100,000-member distribution ($\hat{F}$) will be assumed to be an accurate realisation of the underlying distribution, F. For each distribution, the bootstrapping procedure is repeated 10,000 times in order to remove noise from the sampling distributions of the statistics of interest. Of particular interest is how the uncertainty of these sampling distributions decreases as ensemble size increases. For this purpose, a sampling distribution array of length 10,000 will be created for various ensemble sizes obtained as subsets of the full forecast ensemble. In order to ensure each data point in the distribution had equal weight in the bootstrapping procedure, a jackknife-after-bootstrap analysis was carried out (not shown: Davison and Hinkley, 1997). For what is to follow, it has been determined that no one data point had any significant influence.

For the construction of the confidence intervals, the percentile method is employed, where, for the 95% level,

the 2.5th and 97.5th percentile of the random variable's sampling distribution are the lower and upper bounds to the interval. This is deemed to be appropriate, due to not having knowledge of the underlying distribution and the mostly symmetric nature of the sampling distributions obtained from our bootstrapping procedures.

3 | RESULTS AND DISCUSSION

3.1 | Distributions from the idealised prediction system

The idealised prediction system defined in the previous section reproduces the basic processes of convection and is computationally efficient. The first question to be addressed is whether the forecast distributions generated during the ensemble forecast are representative of a real NWP ensemble system. In this section, distributions will be extracted from the idealised system and their evolution and shape analysed and compared with those of distributions extracted from a 1,000-member NWP ensemble (Craig *et al.*, 2022).

Throughout this study, distributions from the ensemble will be extracted for a single position and time, and as a result will contain 100,000 data points (unless stated). The evolution in time of the shape of the distributions was different, depending on whether the initial condition produced by the DA contained a cloud at the chosen grid point or not. The following subsection therefore will show distributions of the three variables of the idealised model for both initially cloudy and noncloudy grid points.

3.1.1 | Evolution of the wind variable

Figure 4 shows distributions of the wind variable at four time points in the evolution of the free run at initially cloudy and noncloudy grid points. In each histogram, 100 bins are calculated in order to resolve the shapes of the distributions clearly. The histograms are normalised so that the integral is one, with the result that the narrow bin interval leads to probability densities greater than one. The distribution extracted from an initially cloudy grid point shows an increase in spread and tail density until 80 min. At 24 min there is a shift in the mean towards positive

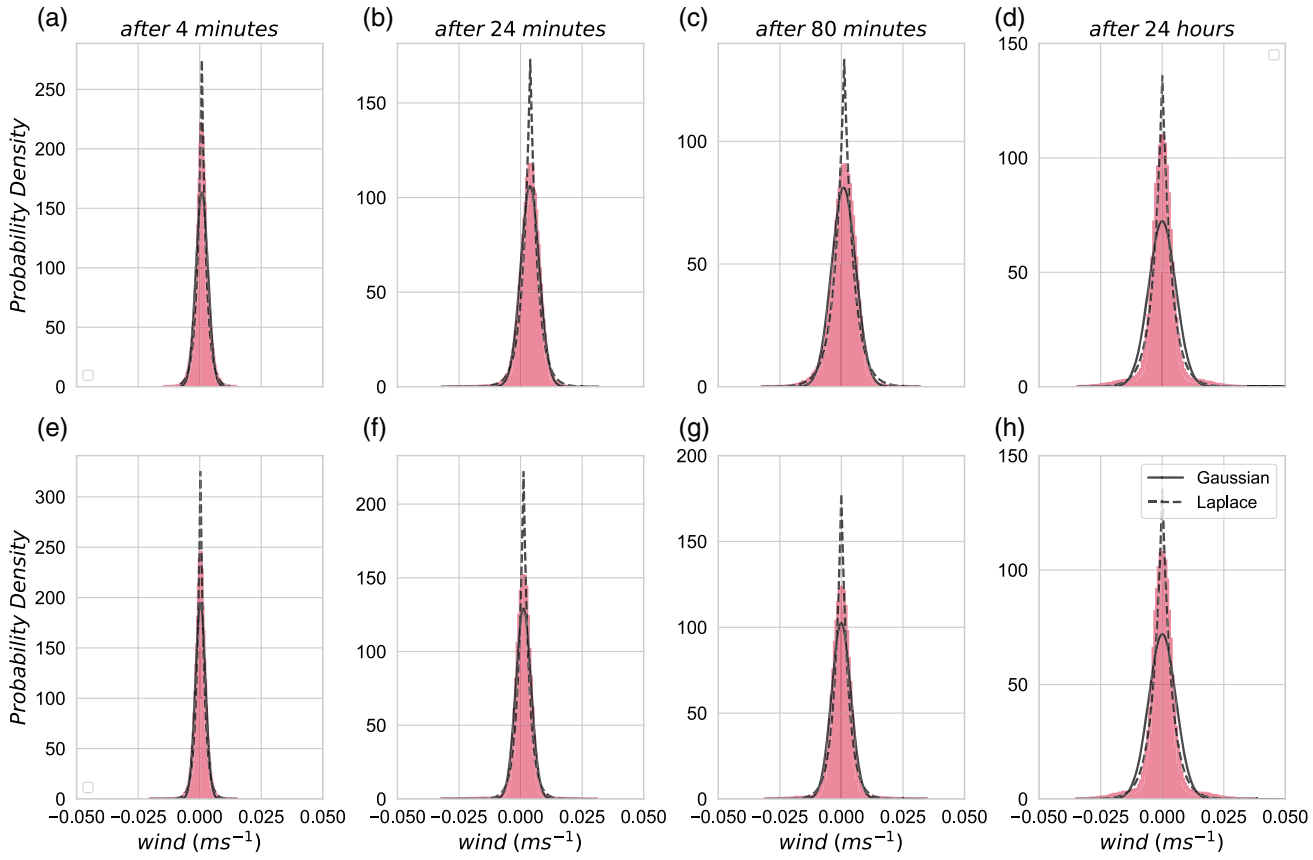

FIGURE 4 Wind variable distributions from the 100,000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 min, and (d,h) at 24 hr of free run, overlaid by a Gaussian and Laplace PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 1 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Non-Gaussian statistics of wind distributions. Table entries are [skewness, kurtosis, KL divergence]

Starting conditions	After: 4 min	24 min	80 min	24 hr
Cloudy	[0.026, 1.309, 0.029]	[−0.374, 1.431, 0.018]	[−0.423, 1.450, 0.024]	[−0.073, 5.356, 0.159]
Noncloudy	[−0.102, 1.789, 0.024]	[−0.539, 7.498, 0.070]	[−0.222, 6.051, 0.078]	[0.016, 5.282, 0.160]

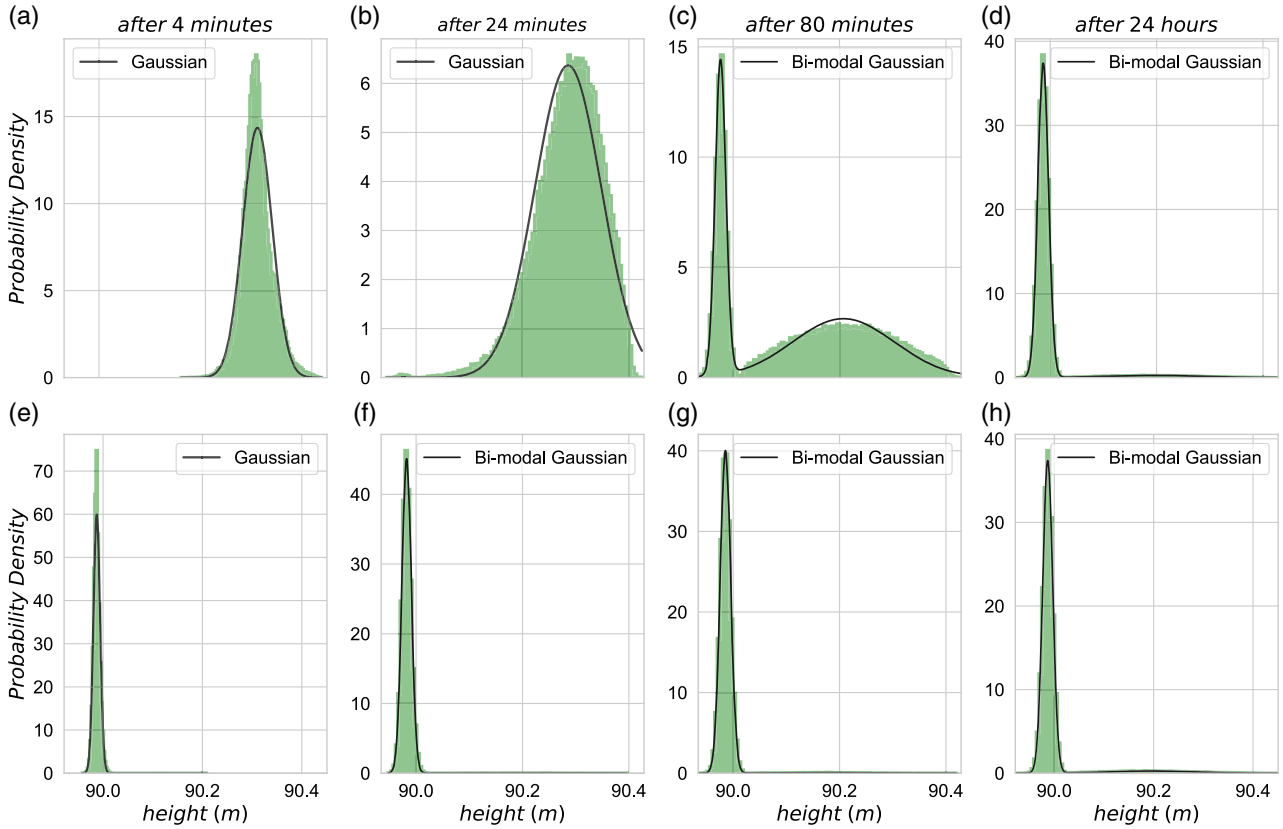

FIGURE 5 Height variable distributions from the 100,000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 min, and (d,h) at 24 hr of free run, overlaid by a Gaussian or bimodal Gaussian PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 2 [Colour figure can be viewed at wileyonlinelibrary.com]

wind, but the mean relaxes gradually to zero again, as seen at 80 min. The distribution at 24 hr is centred around zero. At the initially noncloudy grid point, the distribution follows a similar evolution, except at 24 and 80 min where the mean remains near zero. Table 1 documents three statistics that characterise the non-Gaussianity of the distributions presented in Figure 4. It is clear from the kurtosis that density increases at the tails and the distributions at both grid points become a bit less Gaussian as time evolves. It is interesting to note that the kurtosis of the distribution at the grid point that began with no cloud increases at a faster rate than that at the grid point that started the free run with a cloud. The symmetry of the distributions (small skewness) throughout the evolution is also clear.

At all time points and for both grid points, KL divergence (Table 1) is below 0.3 and as such a Gaussian PDF fits the distributions well. Figure 4 also shows a reference Laplace distribution for comparison. In some cases, the Laplace form can fit aspects of the distribution more effectively than a Gaussian. This is seen at 4 min and at climatology for both grid points where the Laplace form captures the peak of the distribution well. Jacques and Zawadzki (2015) also found their 1,000-member background wind distributions from a convection-resolving forecast to be approximated well by a Laplace PDF.

3.1.2 | Evolution of the height variable

As with the wind variable, the evolving shapes of the height variable distributions (Figure 5) are analysed. The histogram of the height variable at 4 min at the grid point initially containing a cloud shows a single peak with mean above the H_c threshold of 90.02 m. As the ensemble

TABLE 2 Non-Gaussian statistics of height distributions. Entries of table are [skewness, kurtosis, KL Divergence]

Starting conditions	After: 4 min	24 min	80 min	24 hr
Cloudy	[0.352, 1.017, 0.032]	[−0.737, 1.029, 0.052]	[0.244, −1.310, 0.558]	[4.470, 20.828, 1.332]
Noncloudy	[2.206, 43.841, 0.060]	[10.872, 159.423, 0.738]	[8.037, 80.440, 0.896]	[4.426, 20.492, 1.319]

members diverge, some no longer contain a cloud, leading to a second peak which is centred below H_c. This shift can be detected at 24 min, but is clearly visible by 80 min. The formation of a second peak is accompanied by a large increase in KL divergence (Table 2). As time goes on, density in the histogram increasingly shifts to the noncloudy peak (peak with mean below H_c) until the climatological distribution is reached, in which only a few members contain a cloud at that location. The cloudy peak (peak with mean above H_c) is then very small in comparison with the noncloudy peak and the bimodality is hardly visible. The evolution of the distribution for the first 80 min at the grid point that did not initially contain a cloud is roughly opposite to that of the initially cloudy grid point. In this case, the initially noncloudy members gathered below the H_c threshold gradually diverge, with a few members eventually forming clouds to produce a second peak above this threshold.

At 4 min, the distributions at both grid points still show the approximately Gaussian distribution produced by the DA. After 24 and 80 min, a bimodal Gaussian fits the distributions well, for grid points that started both without and with a cloud. This deviation from a simple Gaussian is consistent with the increase in KL divergence to above 0.3 (Table 2) at these grid and time points.

3.1.3 | Evolution of the rain variable

The evolution of the rain variable distributions are shown in Figure 6. Note the log scaled x-axis. The ensemble members below a certain threshold (3×10^{-5}) are considered to have no rain and are not plotted. Instead, the percentage of ensemble members that have rain is stated above each panel. The number of bins is reduced to 50, in order to observe this reduced number of members clearly. In the case of the grid point beginning with a cloud, it contained a cloud that had not yet precipitated. At 4 min, therefore, many of the ensemble members had not yet precipitated. The fraction of members with rain increases up until 80 min, at which time 75% of the members contain rain, compared with 4% at 4 min. Rain is removed by a sink function that is proportional to the rain amount, so that the largest rain amounts experience the most rapid decline, with the result that the peak of the distribution shifts towards smaller values between 24 and 80 min. This is also seen in the strong increase of skewness in Table 3. As the members at 24 hr become decorrelated, there is no well-defined peak as seen at 24 and 80 min. A similar evolution occurs at the grid point that did not initially contain a cloud. However, as the members were not primed to produce rain (they did not already contain a cloud in the updraft phase) fewer members had developed rain at 24 and 80 min. The increase in skewness over the evolution at both grid points is reflected in the divergence from Gaussianity indicated by the KL divergence in Table 3.

When there is significant rain (>0.1% of members), the rain distribution fits a gamma PDF, and, to a lesser extent, a log-normal PDF, well. This distribution shape was also found by Scheuerer and Hamill (2015), where a censored, shifted gamma PDF is fitted in the statistical post-processing of an ensemble reforecast's accumulated precipitation distributions. Note that Figure 6f contains only 17 members with rain; however, it appears it can also be approximated by a gamma/log-normal PDF.

3.1.4 | Comparison with NWP model

Finally, it is important to evaluate whether the form and evolution of the distributions are representative of those found in full NWP systems. The rain and wind speed variables of the idealised model correspond directly with variables of a NWP model, but the height variable requires some interpretation. The most important consideration is that when the height exceeds a certain threshold the buoyancy becomes positive and the grid point is considered to contain a cloud. We therefore identify h with the saturation deficit, or relative humidity, variables that capture the atmospheric variability inside and outside clouds.

For each of the three model variables, the evolution of the distribution shapes has been analysed at a variety of different grid points. It was found that the wind was reasonably well described by a Gaussian or Laplace PDF, height by Gaussian mixture, and rain by a Gamma distribution. This can be compared with the study of 1,000-member ensemble forecasts using an NWP model by Craig *et al.* (2022), where it was found that the distributions of all the forecast variables examined fell into one of three broad categories: quasi-Gaussian, multimodal, or highly skewed. The parameterised fits for the three variables of the idealised model are thus representative

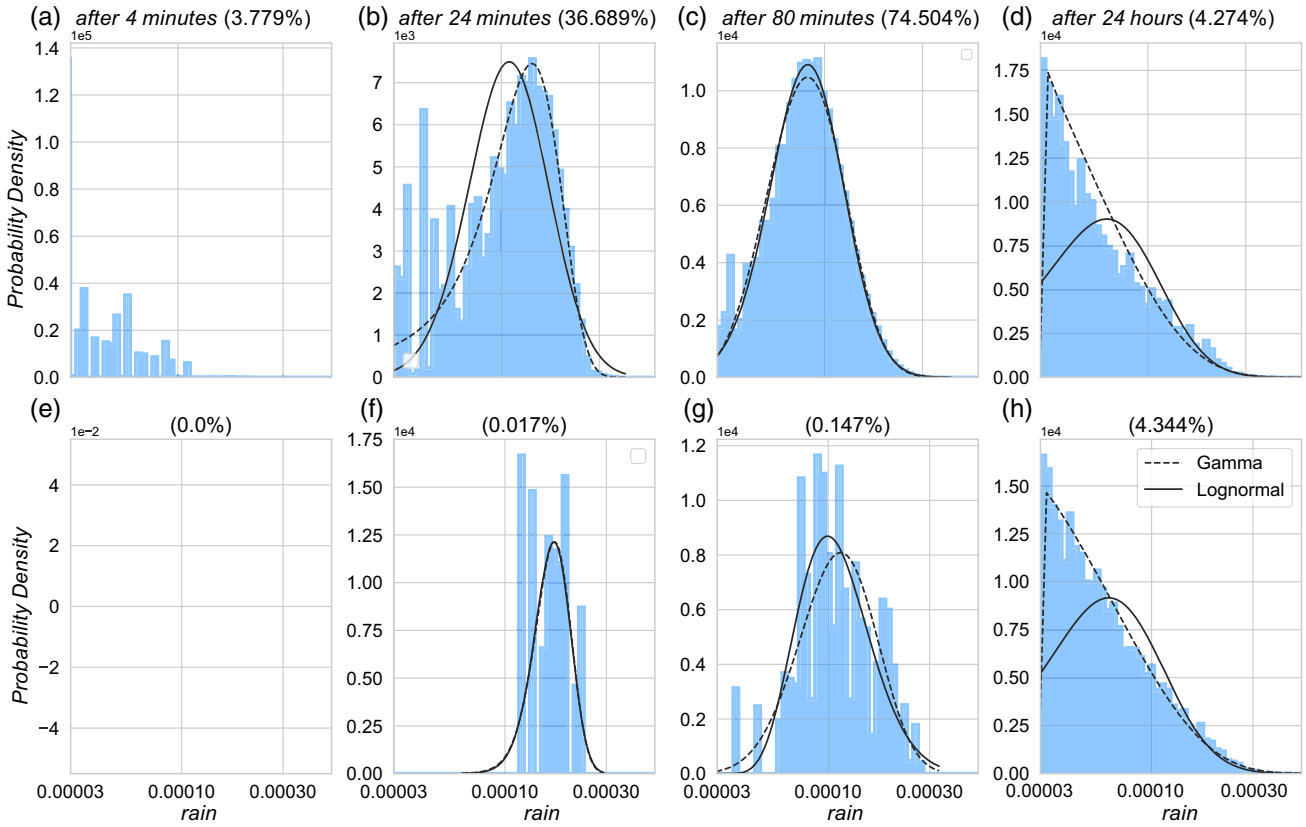

FIGURE 6 Rain variable distributions from the 100,000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 min, and (d,h) at 24 hr of free run. Percentages above histograms show the number of members containing rain, which is shown in the histogram. Overlaid by gamma and lognormal PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 3 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Non-Gaussian statistics of rain distributions. Table entries are [skewness, kurtosis, KL divergence]

Starting conditions	After: 4 min	24 min	80 min	24 hr
Cloudy	[2.164, 7.897, 0.927]	[0.070, −0.285, 0.054]	[1.023, 1.539, 0.097]	[1.388, 3.021, 0.362]
Noncloudy		[0.116, −0.706, 2.033]	[0.511, −0.542, 0.457]	[1.409, 2.536, 0.346]

of the three categories that characterise the NWP ensemble forecasts. Furthermore, the evolution in time of the model variable distributions follows the conceptual model proposed by Craig *et al.* (2022) as described in Section 1. Based on these results, we anticipate that the convergence characteristics of the distributions with ensemble size will also be representative of the behaviour of real-world NWP systems.

A preliminary analysis of the bivariate distributions was carried out in addition. Bivariate distributions were created from pairs of distributions of the same variable, but at different time points, and from pairs of distributions of different variables, but at the same time points. At early time steps of the free run, it was found that bivariate distributions were generally Gaussian, with the exception of those including rain. As time evolved, non-Gaussianity developed as expected, including in those bivariate distributions where both marginal distributions remained Gaussian. This was seen for the case of the bivariate distribution of wind at two different time points, where structures similar to those from the simple model employed by Poterjoy (2022) were created.

3.2 | Sampling uncertainty convergence

The convergence of sampling uncertainty of statistical properties as ensemble size increases is now analysed. Following Craig *et al.* (2022), statistical inference is carried out on selected distributions from the ensemble in the free-run component of the idealised prediction system to identify the nature of the convergence of sampling uncertainty. We

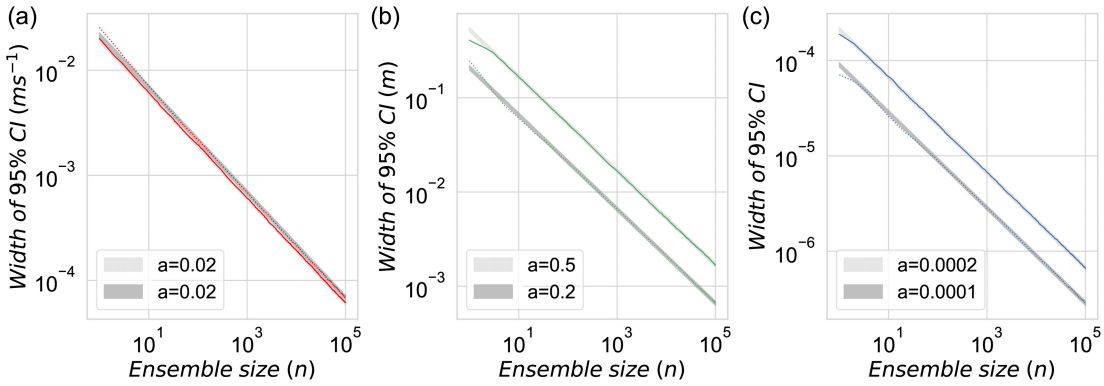

FIGURE 7 Continuous and dotted coloured lines are the width of the 95% confidence interval (CI) of the sampling distribution of the mean for (a) wind, (b) height, and (c) rain model variables. The continuous line uses distributions from Figures 4c, 5c, and 6c. The dotted line uses distributions from Figures 4h, 5h, and 6h. Light and dark grey lines are fitted to continuous and dotted lines respectively, see the text for details. The corresponding width spans 5% above and below the fitted line. The fitted parameter is shown in the legend. The number of ensemble members used for fitting is catalogued in Table A1 in Appendix A [Colour figure can be viewed at wileyonlinelibrary.com]

investigate further how sampling uncertainty convergence is sensitive to the shape of the distribution and the statistic being evaluated.

3.2.1 | Universal convergence scaling characteristic

The analysis of convergence will focus on two cases: the 80-min forecast for an initially cloudy grid point and the 24-hr forecast for an initially noncloudy grid point. As can be seen from panels c and h of Figures 4, 5, and 6, these two cases include the main distribution types found in the forecasts. Note that, for the rain distributions, the zero-rain data points that are omitted from the plots are included in all computations of forecast statistics. For each of the 100,000-member distributions, 10,000 bootstrap distributions were created. Sampling distributions of random variables were then constructed by calculating the desired statistical property for each of the 10,000 bootstrapped distributions. For smaller sample sizes of 1–200 members drawn from the 100,000-member distribution, the random variable sampling distribution of length 10,000 is calculated for every ensemble size. From 200–100,000 members, the random variable sampling distribution is calculated in steps of 100 members. The width of the confidence interval between the 2.5th and 97.5th percentiles of the random variable sampling distribution, which we define as the convergence measure, is subsequently plotted as a function of ensemble size using a log–log scale. The convergence measure is fitted to the expected scaling behaviour of $y = an^{-1/2}$ using linear regression in log space, where a quantifies how the convergence measure scales with n. The range of values used for each fit is detailed in Appendix A. We will describe a forecast statistic as being in the asymptotic regime if the width of the 95% confidence interval of the random variable sampling distribution (the convergence measure) appears to be converging as $n^{-1/2}$ with ensemble size.

The width of the 95% confidence interval of the sampling distribution of the mean, as a function of ensemble size n, is shown in Figure 7 for the three model variables for the two cases. The fitted power-law lines, which scale as $n^{-1/2}$, follow the width of the 95% confidence interval well for each distribution and model variable, except at ensemble sizes below 10 for the height and rain distributions. The decrease of the standard error of the sample mean proportional to $n^{-1/2}$ is an expected result of the CLT. However, the lines corresponding to the two cases are offset from each other, that is, the fitted a values are different. In the case of the mean wind, the difference is small, but for the other variables it is greater than a factor of two. While the asymptotic scaling of the uncertainty appears to be independent of the shape of the underlying distribution, the absolute width of the confidence interval is not. Finally, we note that the convergence measures are similar for rain distributions that both included and did not include zero-rain members (not shown). This is the case for all the results in this study, and for this reason only the convergence measures including the zero-rain members are shown.

The convergence of the sampling distribution for the variance is shown in Figure 8. The power-law scaling of $n^{-1/2}$ is seen again in all distributions. As expected, the CLT is applicable not only to the mean but also to other forecast statistics. The number of members required until convergence appears to follow $n^{-1/2}$ is generally larger than for the mean (Figure 7), and there is an

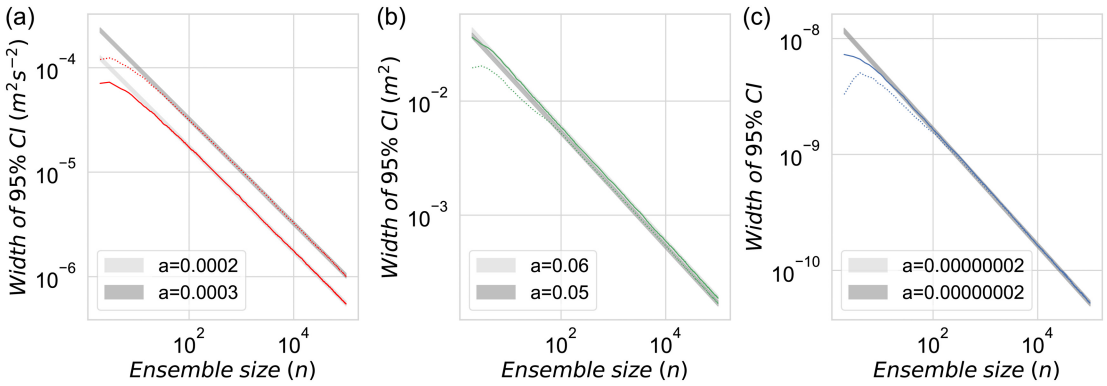

FIGURE 8 As in Figure 7, but for the sampling distribution of the variance [Colour figure can be viewed at wileyonlinelibrary.com]

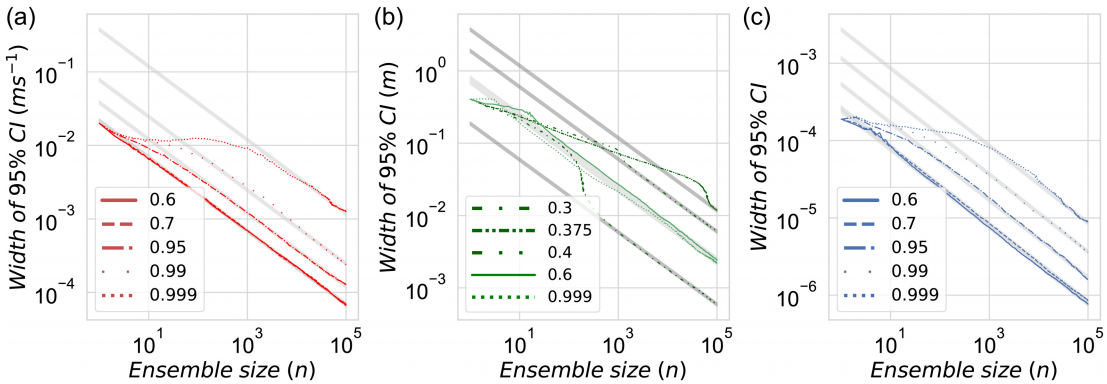

FIGURE 9 As in Figure 7, but for sampling distributions of different quantile levels, p. The legend labels the different quantiles [Colour figure can be viewed at wileyonlinelibrary.com]

overestimation of the width made by the fit at smaller ensemble sizes. This is in line with Craig *et al.* (2022), who discovered that more members are required in the standard deviation compared with the mean in order to achieve convergence as predicted in the asymptotic limit.

The convergence of various quantile sampling distributions is shown in Figure 9. With enough members, it is clear that in most cases the convergence measure scales as $n^{-1/2}$, with wider confidence intervals for more extreme quantiles, as well as more members required to reach the asymptotic regime. This scaling behaviour has also been observed in the skewness and kurtosis (not shown), indicating the universality of the $n^{-1/2}$ scaling of sampling uncertainty with ensemble size as long as enough members are used. The exception was the 0.999 quantile. It could be seen to scale approximately as $n^{-1/2}$, but there was more variability than for the lower quantiles. As such, it is unclear if it has reached the asymptotic convergence regime. Another anomalous behaviour is the apparent downward jump in three of the convergence lines (at $p = 0.3$, 0.375, and 0.4) for the height distribution. We will see in the next section that this is likely due to these quantile levels being situated near the minimum between the two peaks of the height distribution, located at $p = 0.375$. Since these height values are relatively rare, large ensemble sizes are required to provide confident estimates of the distribution shape in this region.

3.2.2 | Dependence on distribution shape

In Section 3.2.1, we found that the convergence measure scales as $n^{-1/2}$ with ensemble size for a sufficiently large ensemble. However, the constant a, and hence the absolute width of the confidence interval, depended on the forecast statistic and on the case being considered. To understand these results better, this section will investigate systematically the effects of the underlying distribution of a forecast variable on the sampling uncertainty for different forecast statistics.

For the wind variable, the distributions for the two cases initially with and without a cloud are very similar (Figure 4c and h). The widths of the confidence intervals for the estimates of the means are also very similar (Figure 7a). When the distribution shapes are less

similar, as for the height and rain distributions in Figures 5 and 6, the differences become substantial. This may be related to the fact that the distribution of the wind variable is near-Gaussian in form, so that the density is greatest near the mean, whereas the multimodal or skewed distributions of the other variables have larger density away from the mean.

The width of the confidence interval for estimates of the ensemble variance also shows differences between the two cases (Figure 8), but for this statistic it is the wind variable for which the difference is largest, while both the height and rain plots show less sensitivity. This may again relate to where the density of the underlying distribution is located, but the connection is less clear.

For confidence intervals of the quantile estimates shown in Figure 9, the majority of convergence lines are displaced from one another. For the unimodal distribution of wind and rain, the further the quantile is from the median, the larger the width of the 95% confidence interval. Hence more uncertainty is attached to those quantiles at the tails compared with those at the centre of the distribution. This behaviour is expected from Equation 1, which states that the standard deviation of a quantile estimate will be inversely proportional to the density of the underlying distribution at the quantile value. The behaviour for the height variable is more complex, with large sampling uncertainties for intermediate quantile values. This is also consistent with Equation 1, however, since the bimodal distribution of h has a minimum near the $p = 0.375$ quantile, leading to wide confidence intervals there.

To show visually the importance of the distribution shape on the convergence of the forecast statistics, contour plots are created showing the ensemble size, n, required to obtain a desired sampling uncertainty for a range of quantiles from a distribution. The values are computed using Equation 1, where the underlying distribution, f, is obtained as a kernel density estimation (KDE) using data from the 100,000-member distribution, using the Scott method to calculate the bandwidth. This leads to the underlying distribution being well represented, but can also lead to the resulting contour lines wavering slightly. Every quantile between 0.01 and 0.99 is calculated in steps of 0.01. Using Equation 1 to estimate a required ensemble size requires knowledge of the underlying distribution. In practice, this must be estimated from an available ensemble, which will typically be much smaller than 100,000 members. For comparison, results will also be shown that are calculated using the bootstrap method employed previously, with three subensemble sizes (50, 100 and 500 members).

Figure 10 shows the resulting contour plot for the near-Gaussian wind distribution (Figure 4c), which resembles the result for a true Gaussian in Figure 1. It can be seen that, for quantiles further away from the median, the number of members required to obtain the same level of uncertainty increases. Similarly, as one moves vertically downwards at a fixed quantile level p, the number of

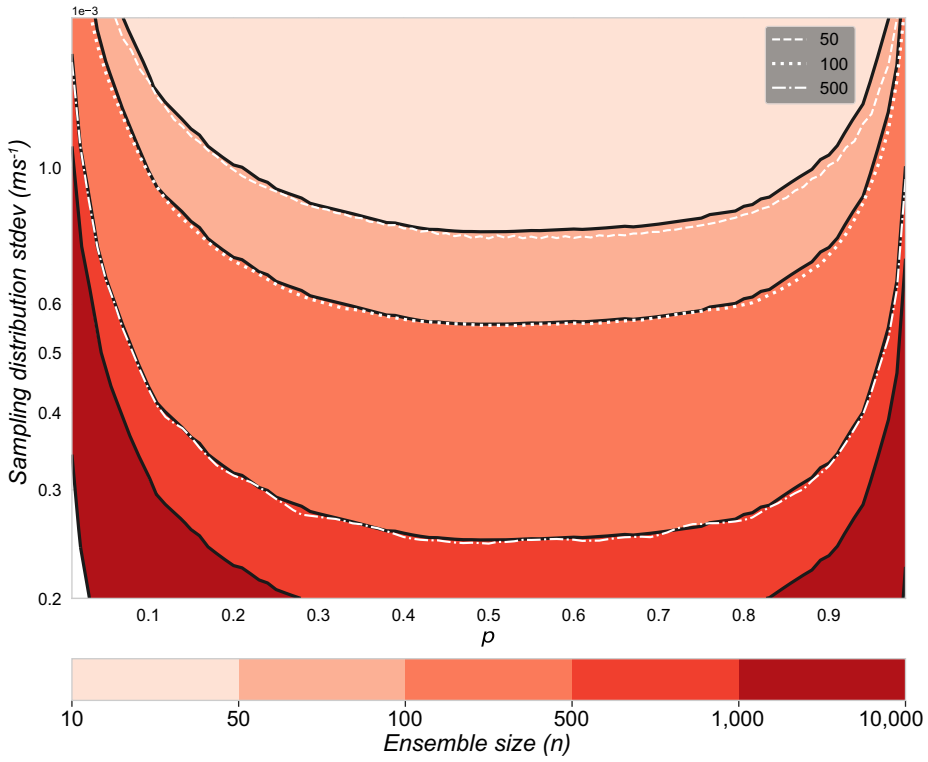

FIGURE 10 Contours show the number of members required to achieve a standard deviation of the quantile sampling distribution according to Equation 1 for quantile levels ranging from 0.01 to 0.99 in steps of 0.01 for the distribution of Figure 4c. White lines show an estimate using the bootstrapping technique [Colour figure can be viewed at wileyonlinelibrary.com]

members required to reach smaller levels of uncertainty increases. As expected, the tails of distributions are more uncertain compared with the peak of the distribution in a unimodal case.

The white lines show estimates obtained with small ensemble sizes. As the number of ensemble members decreases, the estimated value starts to fall below the large ensemble estimate. This is most visible for the 50-member white line. This corresponds to the overestimation of the asymptotic fit in Figure 9a, particularly observable at the 0.95 and 0.99 quantile levels. As the uncertainty calculated in Equation 1 is proportional to $n^{-1/2}$, large deviations between the contours and white lines indicate that the bootstrapped data are not yet converging as $n^{-1/2}$ for that given ensemble size.

As with the wind distribution, a contour plot of n is calculated for the height distribution (Figure 5c) and this is visualised in Figure 11. Unlike for the wind, there is a peak in uncertainty centred around the 0.4 quantile level, which, as noted previously, corresponds to the minimum between the two peaks of the underlying height distribution. This emphasises that any quantile levels corresponding to rare events (such as a trough in the distribution) need more members to obtain the same uncertainty level as at other quantile levels. Since the peak at larger heights (cloudy grid points) is smaller than the other peak, larger ensemble sizes are required for quantiles in this region. A curious feature seen in Figure 11 is the slight decrease in uncertainty in both the large-ensemble and bootstrapped estimates above the 0.96 quantile level. This level corresponds to the rain threshold in Equation 6. Any grid points that surpass this height immediately experience a reduction in buoyancy due to the presence of rain, so that the tail of the distribution is truncated and height values just above this level are not as rare as might be expected. As a result, fewer ensemble members are needed to estimate these quantile levels.

The contour plot of n using the distribution from the rain variable (Figure 6c) is shown in Figure 12. The skewness of the distribution is evident in the asymmetric nature of the contours, with the least uncertain region occurring between p of 0.2 and 0.3 (instead of 0.5). As expected, any p estimate for values outside this region would be more uncertain for the equivalent ensemble size. The longer the tail is, the larger the uncertainty. As the distribution is positively skewed, the quantile levels situated above the peak show larger uncertainties than below. A decrease in uncertainty, analogous to that found for large p in Figure 11, is also seen here, but for quantiles below p of 0.02. As before, this is due to the probability density of f remaining higher than expected, perhaps because the exponential removal of rain leads to an accumulation of rain values close to the zero bound.

3.3 | How big an ensemble do I need?

An important benefit of simple asymptotic scaling for the width of confidence intervals is that an estimation of the number of ensemble members needed to reduce sampling

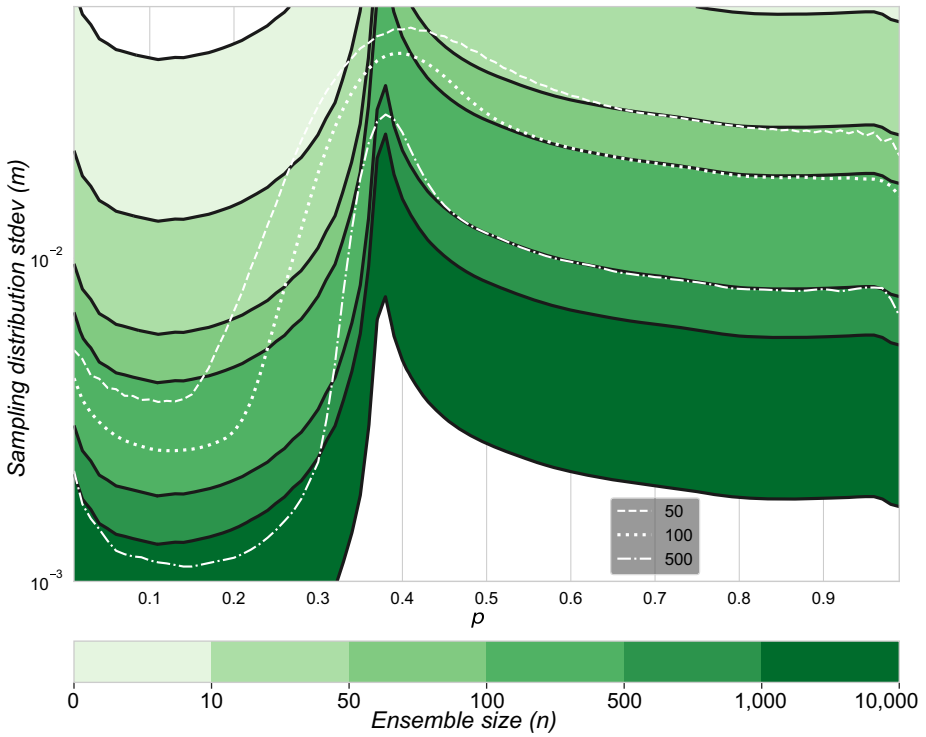

FIGURE 11 As in Figure 10 but with a height distribution from Figure 5c [Colour figure can be viewed at wileyonlinelibrary.com]

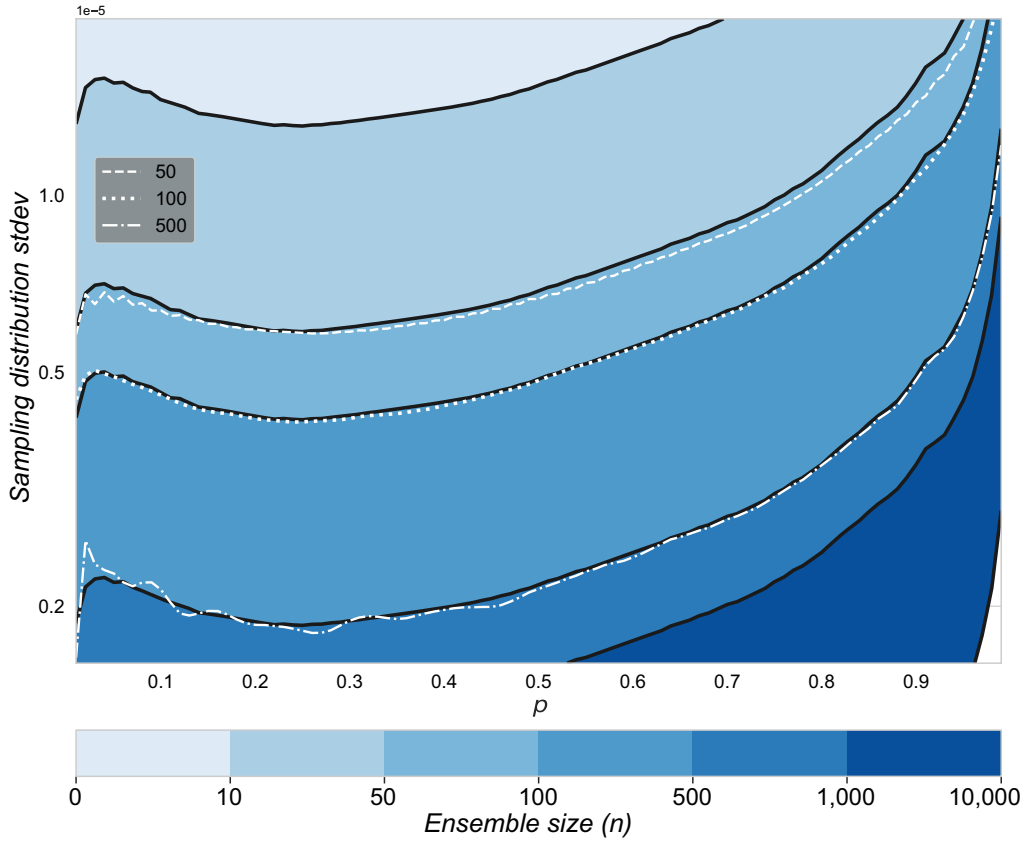

FIGURE 12 As in Figure 10 but with a rain distribution from Figure 6c [Colour figure can be viewed at wileyonlinelibrary.com]

uncertainty to a desired level can be made. This is, of course, only true if the ensemble size is large enough to show that the asymptotic regime is reached. As shown in the previous section, asymptotic convergence could be demonstrated with the 100,000-member idealised ensemble for most statistical properties. It is inconceivable, however, with current computing resources to consider using a 100,000-member NWP ensemble in practice. Hence it is of importance to understand how the asymptotic convergence behaviour may be identified in ensembles of a significantly smaller size. In this section, we will apply two approaches to estimating convergence properties when only small ensembles are available. First, we consider whether asymptotic convergence can be established based on a bootstrap estimate of the uncertainty of the convergence measure from a small ensemble. A second method is then proposed based on a parametric fit of the small ensemble output to an appropriate standard PDF for which the convergence properties can be computed precisely.

3.3.1 | Bootstrapping using smaller ensemble sizes

If only a small ensemble is available for a forecast, it is still possible to construct a bootstrap estimate of confidence intervals as before, but these estimates may not be useful if the small ensemble is not representative of the full distribution. To investigate this issue, we first construct confidence intervals based on different small ensembles drawn from the 100,000 members computed previously, to see whether the convergence behaviour is consistent. Figure 13 shows convergence curves for a sample of forecast variables, namely the variance and selected quantiles of the wind, height, and rain distributions (see Figures 4c, 5c, and 6c respectively). This includes variables that converge for relatively small ensemble sizes, as well as more extreme values that occur only rarely. The plots show convergence curves computed by bootstrapping from ensembles of size 50, 100, 500 and 1,000. Each calculation is repeated 10 times for different small ensembles of the given size. For reference, the curves constructed from the 100,000-member ensemble are also plotted.

For the variables on the top row of Figure 13, even 50 members is sufficient to identify the asymptotic convergence regime, with the width of the confidence interval scaling as $n^{-1/2}$. It is interesting that the correct scaling behaviour is found for the estimates based on smaller ensemble sizes, although there is spread in the curves, which generally increases as the ensemble becomes smaller and there is an offset from the 100,000 member

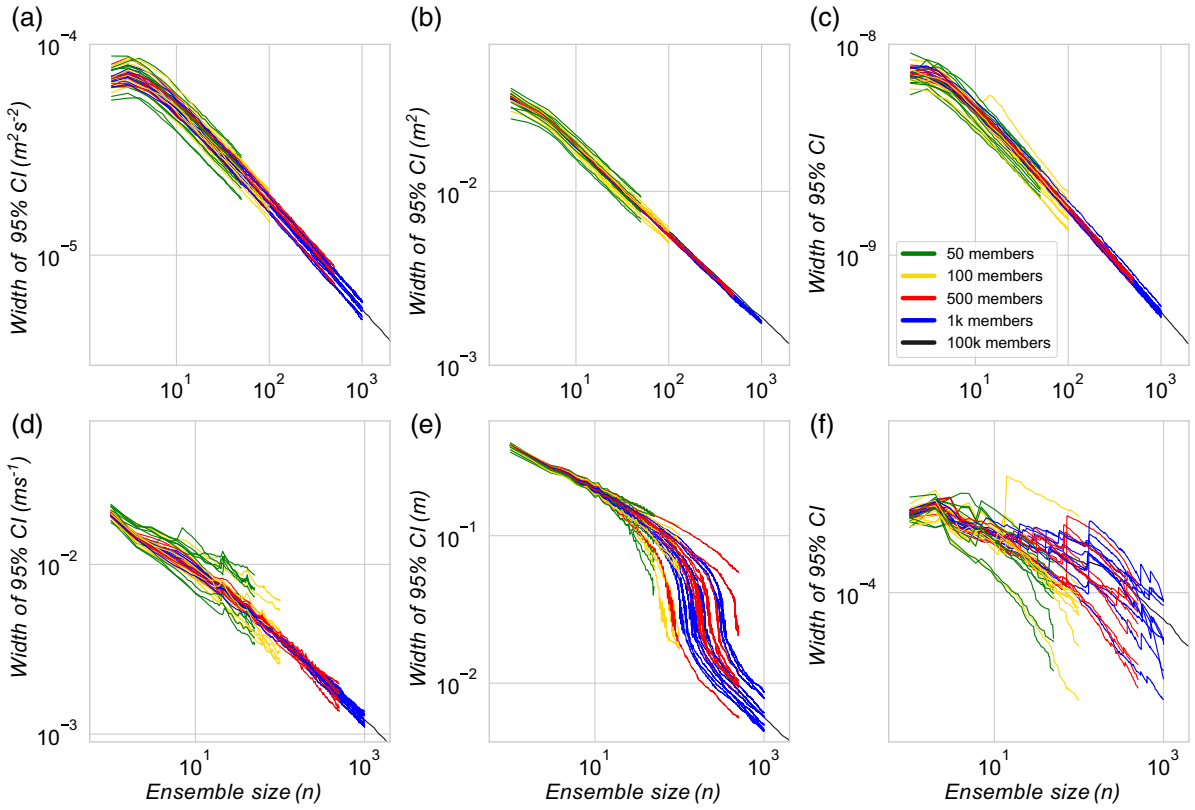

FIGURE 13 Width of 95% CI of the sampling distribution of (a) variance and (d) 95th percentile of wind distribution (Figure 4c), (b) variance and (e) 30th percentile of height distribution (Figure 5c), and (c) variance and (f) 99.9th percentile of rain distribution (Figure 6c) as a function of ensemble size. Convergence measures are calculated 10 times using different sizes of ensemble (50, 100, 500 and 1,000 members), which are different samples of the full 100,000 distribution. The convergence measure calculated using all 100,000 members is in black in the background [Colour figure can be viewed at wileyonlinelibrary.com]

black line. Figure 13d shows an example in which asymptotic scaling is seen only for estimates based on ensemble sizes of 500 members or larger. The curves based on smaller ensemble sizes show a range of slopes, giving a clear indication that the ensemble is not large enough to show convergence behaviour. Note, however, that, while it is unlikely, it is not impossible to find a small ensemble that gives the $n^{-1/2}$ slope by chance. Figure 13e shows the interesting case of the 30th percentile of the height distribution, near the minimum between the two peaks. As noted earlier, small ensembles do not have sufficient resolution to distinguish the peaks, and show asymptotic behaviour for a limited range of n before dropping to the true convergence curve when n becomes sufficiently large. The curves based on small ensembles all follow this behaviour, but if the ensemble is not large enough to resolve the two peaks of the height distribution, it will appear as though the asymptotic regime has been reached. Finally, the extreme rain example in Figure 13f shows no evidence of convergence for any of the ensemble sizes considered here.

The previous figure shows that if an ensemble is large enough to be in the asymptotic convergence regime for a forecast variable, the scaling behaviour will be seen in plots of the confidence interval, but with a random offset that would affect the accuracy of an extrapolation of the results to large ensemble sizes. If the ensemble is not large enough to show asymptotic convergence, the results show a large variability among different realisations of the small ensemble. In practice, this variability will not be seen, because only a single realisation of the ensemble will be available. However, multiple realisations can be generated by bootstrap resampling, and we pose the question of whether a set of ensembles generated this way shows the same variability as an ensemble drawn from the full distribution.

Figure 14 investigates this for the case of a 50-member ensemble. For reference, blue lines show convergence curves for 10 ensembles drawn from the 100,000 member data set. These are overplotted with 100 curves generated from 100 ensembles generated by resampling a single 50-member ensemble. For most of the forecast variables, the resultant spread (red lines) is similar to that of the blue lines. This is the case for all the variance and extreme quantile measures, except for a slight overestimation of uncertainty of the 30th percentile of height. This suggests that it will often be possible to determine whether a given ensemble forecast is large enough to produce asymptotic scaling

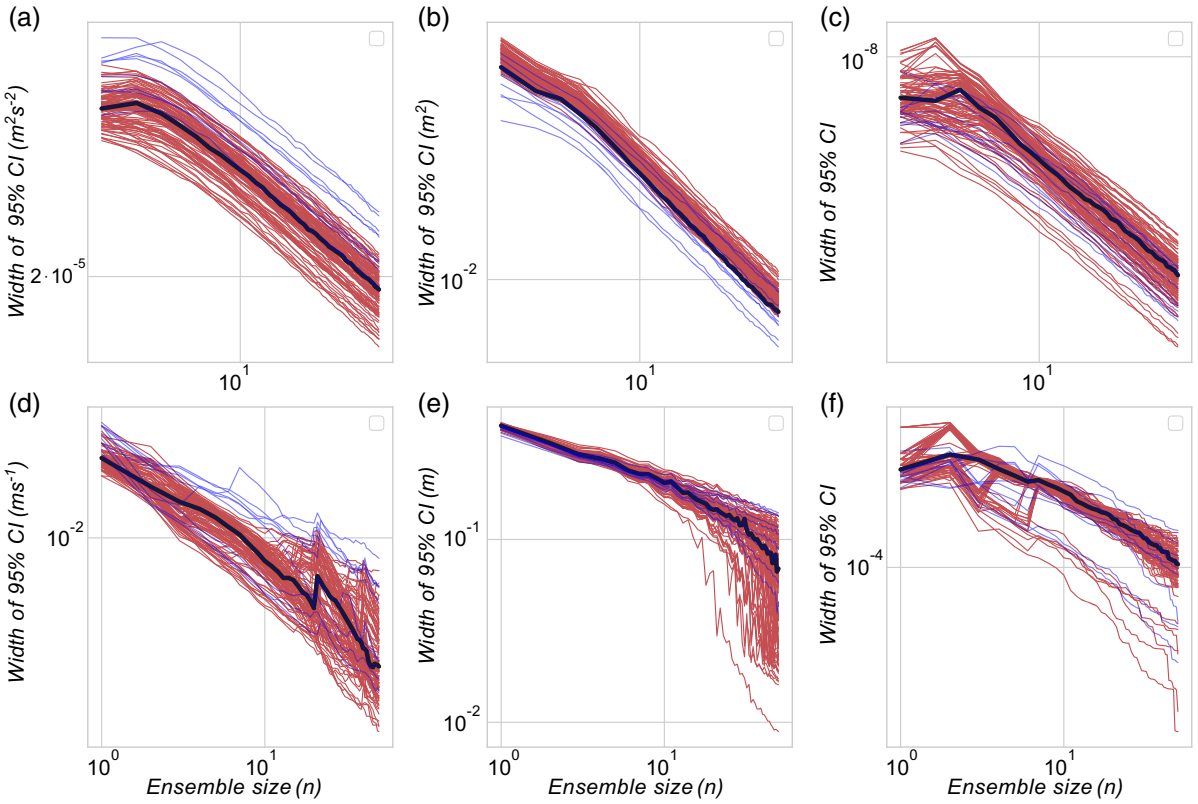

FIGURE 14 (a–f) As in Figure 13. Blue lines are green lines from Figure 13. One sampled distribution of size 50 from the original 100,000-member ensemble was bootstrapped to obtain 100 distribution samples of size 50. The convergence measure calculated from these distributions is shown in red. The sampled distribution used for red lines has its own convergence measure shown in black [Colour figure can be viewed at wileyonlinelibrary.com]

behaviour. If this is the case, the estimate of the sampling uncertainty can reliably be extrapolated to predict how uncertainty will decrease with ensemble size.

3.3.2 | Parameterisation of distributions

As we have seen, it is possible to determine whether the sampling uncertainty of a statistical property of an ensemble's prognostic variable is converging asymptotically or not. For many quantities of interest, however, especially extreme events, the conclusion will be that the ensemble is too small and the estimates of sampling uncertainty will not be reliable. In this section, we explore the potential of using a priori knowledge of the distribution of a forecast variable to provide improved estimates from such small ensembles. Figures 4,5, and 6 showed how distributions from the free run of the idealised prediction system can be classified into three categories. It is then possible to estimate using a small number of members how the distribution with a much larger ensemble would look by assuming one of these three categories as the underlying PDF. The convergence measure can then be calculated using a smaller ensemble. For example, in the case of a Gaussian fit, the mean and standard deviation parameters would be calculated from the data. With this fitted Gaussian, a dataset of members of any size could be generated. This dataset could then be used to calculate the convergence measure using the bootstrapping method as before. In the following, both the full ensemble and 50 members from the 100,000-member ensemble are used to create parameterised distributions, the resulting convergence of which will be compared. From the results, we can conclude whether the parameterisation technique can calculate the convergence measure accurately, and how accurate it is when only 50 members are used.

The convergence measure of the mean calculated using distributions generated from a parametric fit of a wind and height distribution (Figures 4c and 5c, respectively) is shown in Figure 15. The parameterisations (Gaussian for wind and bimodal Gaussian for height), which used two different sizes of ensemble for parameterisation (each shown by grey and black lines), showed good agreement with the convergence calculated using the original 100,000-member ensemble data.

The convergence measure of the variance could be estimated approximately by parameterisation of the ensemble distribution (Figure 16). The convergence measure calculated with the two parameterisations, however, is displaced for both distributions. In the case of the wind distribution, the parameterisation creates an underlying PDF that has smaller variance. This leads to the resulting sampling distribution of variance being smaller and hence produces a narrower 95% confidence interval. This is also the reason for the shift in the case of the height distribution. Note that using more than 50 members for the parameterisation does not improve the results greatly.

The use of parameterisation to estimate the sampling uncertainty of quantiles is now investigated. In Figure 10, it was found that, when estimating f with a KDE using the full 100,000-member ensemble, Equation 1 gave a good approximation to the bootstrapped measurements of convergence, indicating that the convergence of uncertainty was described well by asymptotic theory. This was generally also the case for the height and rain model variables. The black contours of Figure 17a show the number of ensemble members required for a certain standard deviation of the quantile sampling distribution for a range of quantiles as before, but now calculated using a Gaussian parameterisation for f. We use 50 members from Figure 4c to estimate the Gaussian parameters. Although the parameterisation estimate of the convergence measure seems relatively accurate, the Gaussian is not fitted precisely to the KDE (grey lines), which was estimated with 100,000 members. There is an underestimation of uncertainty below p of 0.3 due to the KDE density being smaller than the Gaussian density in this region. When the KDE is estimated using 50 members from the ensemble for f, it gives an imprecise estimate of the uncertainty (purple lines). The difference in accuracy between the KDE estimated from 50 members and the parameterisation when 50 members are used is clear. At small ensemble sizes, the parameterisation method has a much greater accuracy than using KDE for estimating f in Equation 1. This is also the case for the height and rain distributions discussed below (not shown). When 100,000 members are used to fit the Gaussian (contours of Figure 17b), there is a lesser underestimation of uncertainty below p of 0.3. However, 50 members generally gives closer alignment to the KDE than estimation with 100,000 members.

A bimodal Gaussian parameterisation of the height distribution shown in Figure 5c is used to estimate the convergence of sampling uncertainty in the quantiles (Figure 18). Unlike nonparametric methods, the fitted bimodal distribution always produces a qualitatively correct structure, but, when only 50 members are used for the fit, the p value at which the transition between the two peaks occurs is displaced by about 0.15 and it is no longer a good estimation of the convergence measure. When 100,000 members are used to parameterise, the uncertainty estimate is closer to the KDE using 100,000 members for its estimation, but with slight over- and underestimation of uncertainty in

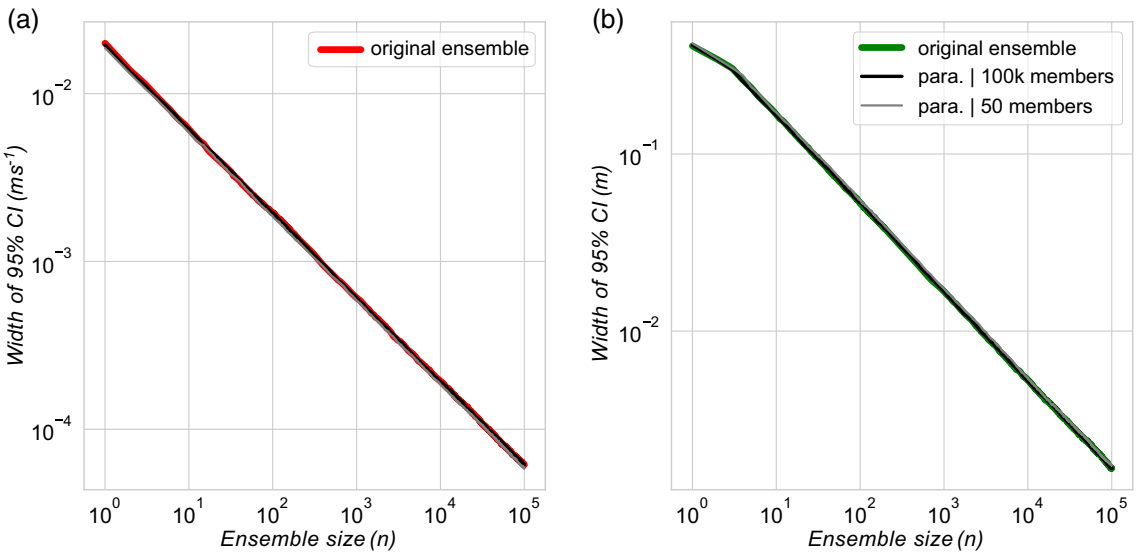

FIGURE 15 (a) Red and (b) green lines are width of 95% CI of sampling distribution of mean for (a) wind distribution (Figure 4c) and (b) height distribution (Figure 5c), calculated using bootstrapping using the 100,000-member ensemble data. Black lines show convergence using data generated from a fitted parameterisation that used 100,000 members from the ensemble. Grey lines show convergence using data generated from a fitted parameterisation that used 50 members from the ensemble [Colour figure can be viewed at wileyonlinelibrary.com]

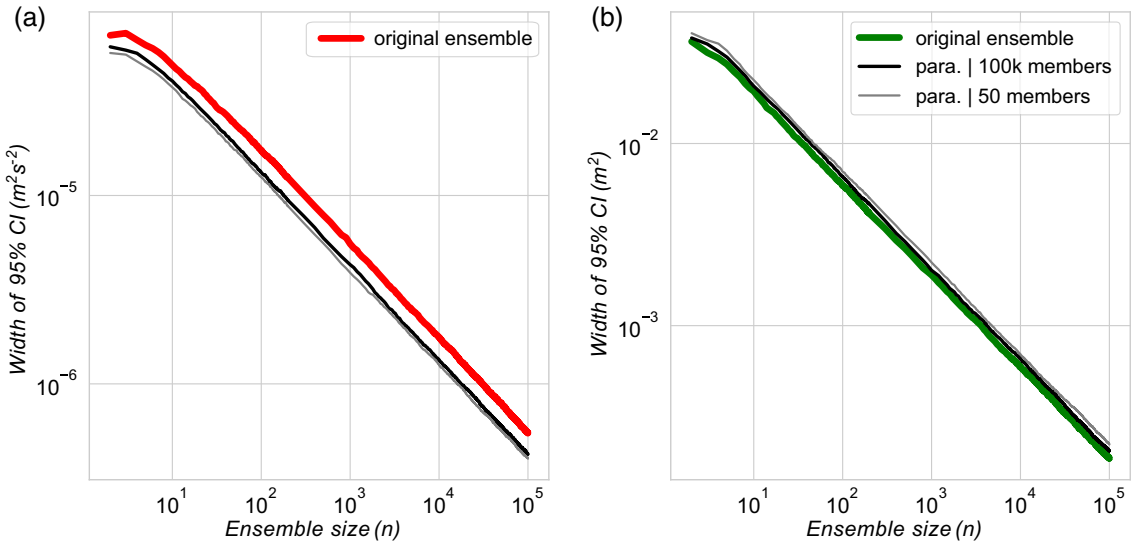

FIGURE 16 As in Figure 15, but for the sampling distribution of the variance [Colour figure can be viewed at wileyonlinelibrary.com]

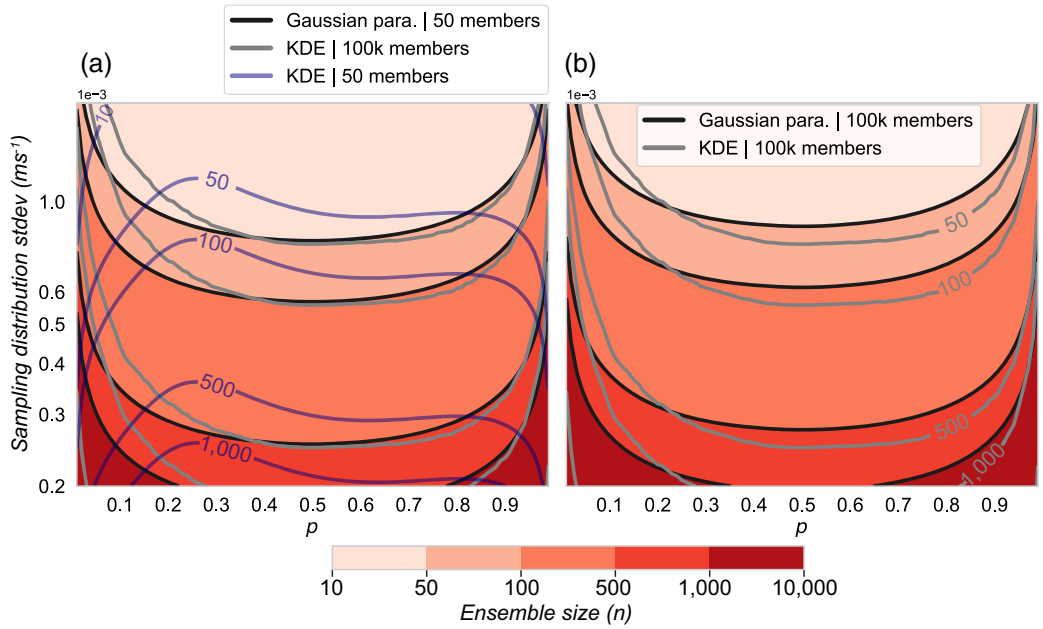

FIGURE 17 Contours created as in Figure 10, but using (a) 50 and (b) 100,000 members from the distribution of Figure 4c to parameterise f. The grey line shows the outline of the contour of Figure 10. The purple line (a) is the result from using a KDE estimated using only 50 members for f [Colour figure can be viewed at wileyonlinelibrary.com]

regions. Only the bimodal Gaussian parameterisation calculated using 50 members from the ensemble captures the decrease in uncertainty above the 0.96 quantile level.

Parameterising the rain distribution of Figure 6c with a Gamma PDF results in reasonable uncertainty estimates of the convergence of the sampling uncertainty of quantiles (Figure 19). In the two uncertainty estimates from each of the parameterisations, there is a slight underestimation at small p values below 0.2. This underestimation is larger, and occurs for a larger range of quantiles, for the parameterisation that used only 50 members. The underestimation occurs due to the difference in the density of the tails of the KDE and the parameterised distributions. The decrease in uncertainty below the 0.02 quantile level is not captured by either method.

From Figures 15–19, it is clear that using a relatively small number of members to calculate the convergence measure by using a parameterised distribution can be reasonably accurate. It has been found that 50 members are

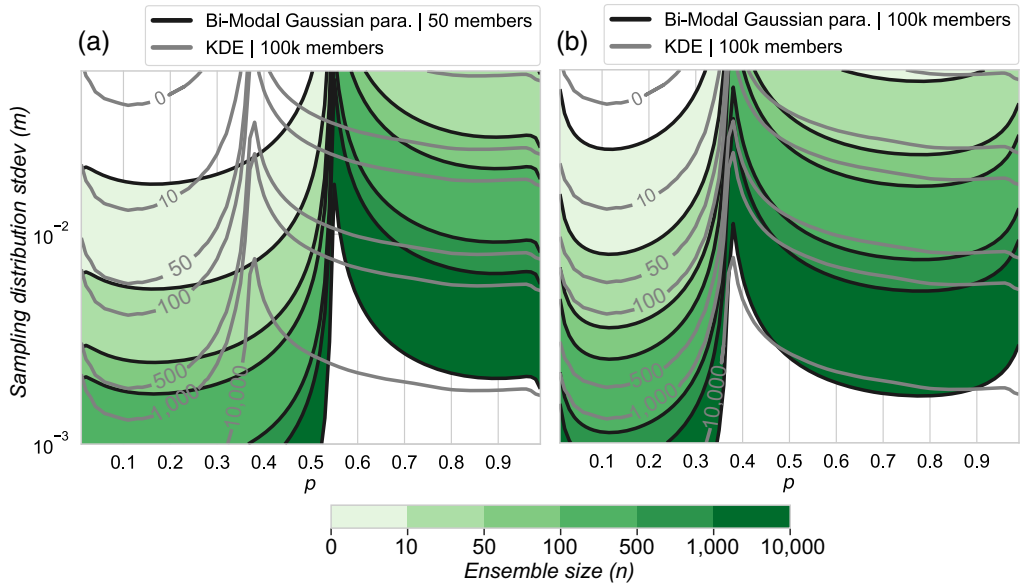

FIGURE 18 Contours created as in Figure 11, but using (a) 50 and (b) 100,000 members from the distribution of Figure 5c to parameterise f. The grey line shows the outline of the contour of Figure 11 [Colour figure can be viewed at wileyonlinelibrary.com]

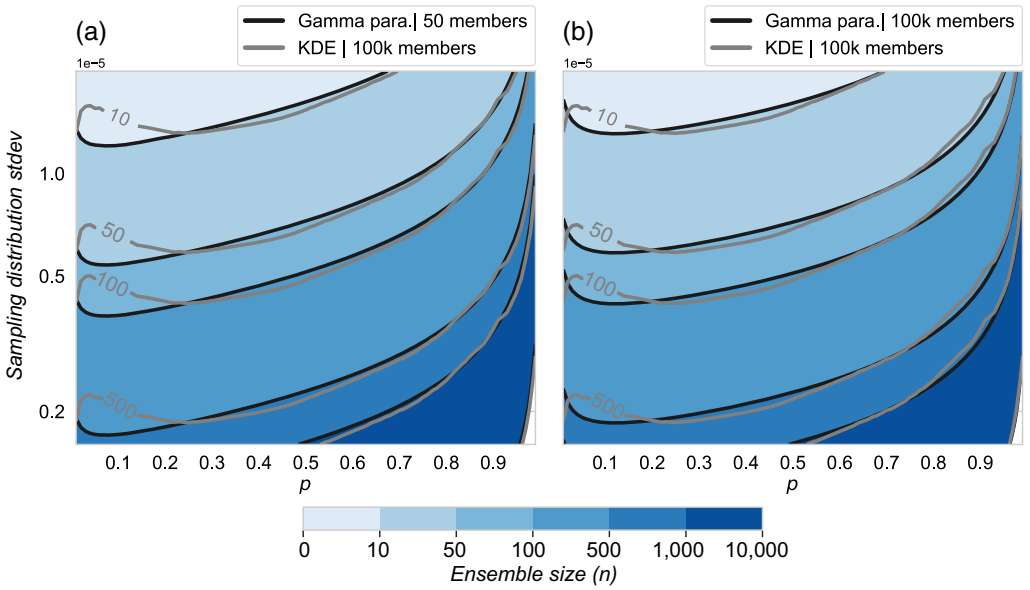

FIGURE 19 Contours created as in Figure 12, but using (a) 50 and (b) 100,000 members from the distribution of Figure 6c to parameterise f. The grey line shows the outline of the contour of Figure 12 [Colour figure can be viewed at wileyonlinelibrary.com]

enough to estimate the convergence measure of the mean and variance, as well as quantiles of "simple" unimodal distributions. More members would be required for distributions with a multimodal shape. It has been seen that there is little benefit in using a KDE approximation to the full distribution with a small number of ensemble members to estimate the uncertainty of quantiles. As there is no limit to how many members can be generated from a parametric fit, this method can be used to obtain the characteristics of the asymptotic convergence as long as the shape of the underlying distribution is captured.

4 | CONCLUSIONS

Operational probabilistic forecasting is limited to relatively small ensemble sizes due to high computational costs. This can impact how representative of the truth the underlying distribution is by creating a sampling uncertainty. While the sampling uncertainty is expected to decrease with increasing ensemble size, it is difficult to determine what ensemble size is required to reduce it to a desired level. In this study we used an idealised prediction system, which replicates the key processes of

convection, to identify how sampling uncertainty of statistical properties converges with ensemble size and to assess the relevance of asymptotic theory to ensemble weather prediction.

The one-dimensional idealised prediction system employed in this study was evaluated by comparing the ensemble distribution shapes for the three prognostic variables with corresponding quantities from a 1,000-member ensemble based on a full convection-permitting numerical model (Craig *et al.*, 2022). Shapes of these distributions over a 24-hr evolution in the free run were found to fit into three categories: quasi-Gaussian, multimodal, and highly skewed, as in Craig *et al.* (2022). As expected from previous work, the distributions became less Gaussian-distributed in time, as expected due to nonlinear convective processes (Zhang, 2005; Legrand *et al.*, 2016; Kondo and Miyoshi, 2019; Kawabata and Ueno, 2020; Craig *et al.*, 2022; Poterjoy, 2022).

In the limit of large n, the sampling uncertainty (width of confidence interval) was found to scale universally as $n^{-1/2}$. This applied to statistical properties including the mean, variance, quantiles between 0.01 and 0.99 as well as 0.999, skewness, and kurtosis. The point at which asymptotic convergence is reached, and the magnitude of the sampling uncertainty, depends on the statistical quantity and the distribution shape. In general, the more the statistic depends on extreme or infrequent values, the more members are required to reach convergence. Since this behaviour does not depend on the distributions being Gaussian, this conclusion should continue to hold for multivariate distributions, where non-Gaussianity is often stronger than for marginals. However, due to the larger uncertainty associated with quantities from multivariate distributions, we expect that the absolute level of uncertainty (width of confidence intervals) would be larger than for their unimodal counterparts. We also expect that more members would be required for asymptotic convergence to be reached.

For the quantiles, the dependence of sampling uncertainty on distribution shape could be described by Equation 1, which states that the sampling uncertainty is inversely proportional to the frequency of occurrence of a quantile. The applicability of this equation to the simulated large ensemble distributions highlights the relevance of asymptotic theory to ensemble weather prediction. This observed theory can be used to provide an alternative method to estimate how adding ensemble members would improve a probabilistic forecast and, in extension, to determine how large an ensemble should be. This way of thinking contrasts with studies such as Leith (1974), which provides a specific number of members required to achieve sufficient precision in a specific aspect of an ensemble. Rather, the asymptotic convergence provides a scaling rule that can be used to answer the question of how large an ensemble should be based on individual ensemble requirements, provided the ensemble is sufficiently large for the theory to apply.

The question of how to apply the asymptotic theory to small ensembles, where it is not obvious that the large n theory is applicable, was addressed in two ways. First, the uncertainty of the convergence measure could be used to determine whether asymptotic convergence had already been reached. If this was not the case, parameterisation of the underlying distribution could be employed. In this case, a good estimate of the convergence measure could be calculated if an appropriate form for the distribution shape was assumed. In an operational setting, the underlying distributions could potentially be obtained from reforecasts.

The ability to quantify the convergence of sampling uncertainty of statistical quantities in ensembles of operational size allows us to address the question of how many ensemble members are needed. For example, an operational forecaster would like to know whether it would be worthwhile investing in expanding the current 50-member NWP ensemble to 100 members, and is particularly interested in the ability to estimate the spread of temperature over Munich accurately. To answer this question, the forecaster would like to calculate the convergence measure of the variance statistic for the temperature variable. The first thing required is to check whether asymptotic theory can be applied, by calculating the uncertainty in the convergence measure. This is done by bootstrapping the 50-member ensemble 100 times to obtain 100 distributions of length 50. With each of these distributions, the convergence measure is then calculated. The forecaster finds no divergence in the measures, similar to the green lines of Figure 13a, that is, the convergence is in the asymptotic regime. This enables visualisation of how the 95% confidence interval width will decrease as extra ensemble members are added to the 50-member NWP ensemble and hence how the accuracy of the estimate of the range of temperature over Munich will increase as more members are added. Knowledge of how many members to aim for in the future to obtain a certain level of sampling uncertainty can hence be calculated.

The idealised prediction system developed in this study does not contain the complexity of a full NWP system. This made it possible to create a huge ensemble, which allowed us to look extensively at the convergence behaviour of the sampling uncertainty. Many physical processes and sources of error in the atmosphere are, however, not represented in the idealised system. Therefore results from more complex systems are vital to have, in combination with those from this study (e.g. Craig *et al.*, 2022). One missing aspect is the dependence on weather regime, particularly

the influence of weak and strong forcing of convection (Keil *et al.*, 2014; 2020). Furthermore, this study did not consider techniques to inflate the effective ensemble size, such as the neighbourhood method, and how they may affect the convergence behaviour (Ebert, 2009; Ben Bouallègue *et al.*, 2013; Hagelin *et al.*, 2017). Finally, a homogeneous domain and boundary conditions were used, which resulted in all grid points following a similar evolution in their distribution shapes, with the only important distinction being whether they started the simulation with or without a cloud. An important example would be the effects of orography (Bachmann *et al.*, 2020). These are areas for future research.

A final caveat is that the method here considers only sampling uncertainty and its dependence on ensemble size. Other sources of error in ensemble predictions, including model error and initial-condition error resulting from limited observations or approximations in the DA system, will limit the accuracy of probabilistic forecasts regardless of ensemble size.

AUTHOR CONTRIBUTIONS

Kirsten I. Tempest: data curation; formal analysis; investigation; methodology; visualisation; writing – original draft; writing – review and editing. **George C. Craig:** conceptualisation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; supervision; visualisation; writing – review and editing. **Jonas R. Brehmer:** formal analysis; investigation; methodology.

ACKNOWLEDGEMENTS

The research leading to these results has been carried out within subproject A6 of the Transregional Collaborative Research Center SFB/TRR 165 "Waves to Weather" (https://www.wavestoweather.de/) funded by the German Research Foundation (DFG). J.R.B. is, in addition, grateful for funding by the Klaus Tschira Foundation and for infrastructural support provided by the University of Mannheim. Open Access funding enabled and organised by Projekt DEAL.

ORCID

Kirsten I. Tempest https://orcid.org/0000-0002-2318-9032
George C. Craig https://orcid.org/0000-0002-7431-8164
Jonas R. Brehmer https://orcid.org/0000-0002-8867-707X

TWITTER

Kirsten I. Tempest @kirsten_tempest

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. URL: https://doi.org/10.1175/MWR-D-19-0045.1

Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55. URL: https://doi.org/10.1038/nature14956

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

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. URL: https://journals.ametsoc.org/view/journals/mwre/140/11/mwr-d-12-00031.1.xml

Buizza, R., Barkmeijer, J., Palmer, T.N. and Richardson, D.S. (2000) Current status and future developments of the ecmwf ensemble prediction system. *Meteorological Applications*, 7, 163–175. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482700001456

Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A. and Wedi, N. (1998) Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 124, 1935–1960. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712455008

Cohen, B.G. and Craig, G.C. (2006) Fluctuations in an equilibrium convective ensemble. part II: Numerical experiments. *Journal of the Atmospheric Sciences*, 63, 2005–2015. URL: https://journals.ametsoc.org/view/journals/atsc/63/8/jas3710.1.xml

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

Craig, G.C. and Cohen, B.G. (2006) Fluctuations in an equilibrium convective ensemble. part I: Theoretical formulation. *Journal of the Atmospheric Sciences*, 63, 1996–2004. URL: https://journals.ametsoc.org/view/journals/atsc/63/8/jas3709.1.xml

Davison, A. and Hinkley, D. (1997) *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press.

Dekking, F., Kraaikamp, C., Lopuhaä, H. and Meester, L. (2005) *A Modern Introduction to Probability and Statistics*. Springer.

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

Evensen, G. (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99, 10143–10162. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572

Feng, X., DelSole, T. and Houser, P. (2011) Bootstrap estimated seasonal potential predictability of global temperature and precipitation. *Geophysical Research Letters*, 38. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2010GL046511

Gaspari, G. and Cohn, S.E. (1999) Construction of correlation functions in two and three dimensions. *Quarterly Journal of the*

Royal Meteorological Society, 125, 723–757. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555417

Gneiting, T. (2014) *Calibration of Medium-Range Weather Forecasts*. URL: https://www.ecmwf.int/node/9607.

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. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3135

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. URL: https://doi.org/10.1175/MWR-D-19-0060.1

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. URL: https://journals.ametsoc.org/view/journals/mwre/143/7/mwr-d-14-00243.1.xml

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. URL: https://journals.ametsoc.org/view/journals/mwre/145/4/mwr-d-16-0160.1.xml

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. URL: https://journals.ametsoc.org/view/journals/mwre/148/1/mwr-d-18-0367.1.xml

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. URL: 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. *Q.J.R. 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://npg.copernicus.org/articles/26/211/2019/

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

Leith, C. (1974) Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102, 409

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

Leutbecher, M. and Palmer, T. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227, 3515–3539. Predicting weather, climate and extreme events. https://www.sciencedirect.com/science/article/pii/S0021999107000812

Lin, J., Emanuel, K. and Vigh, J. (2020) Forecasts of hurricanes using large-ensemble outputs. *Weather and Forecasting*, 5, 1713–1731.

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

Met Office (2022) *The Met Office ensemble system*. https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/mogreps. [Accessed 24th September 2022].

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. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL060863

Poterjoy, J. (2022) Implications of multivariate non-gaussian data assimilation for multiscale weather prediction. *Monthly Weather Review*, 150, 1475–1493. URL: https://journals.ametsoc.org/view/journals/mwre/150/6/MWR-D-21-0228.1.xml

Rasp, S., Selz, T. and Craig, G.C. (2017) 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 the Atmospheric Sciences*, 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. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3159

Reinert, D., Prill, F., Frank, H., Denhard, M., Baldauf, M., Schraff, C., Gebhardt, C., Marsigli, C. and Zängl, G. (2020). *DWD database reference for the global and regional icon and icon-eps forecasting system*. Tech. rep., Technical Report Version 2.1. 1.

Richardson, D.S. (2001) Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127, 2473–2489. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712757715

Ruckstuhl, Y., Janjić, T. and Rasp, S. (2021) Training a convolutional neural network to conserve mass in data assimilation. *Nonlinear Processes in Geophysics*, 28, 111–119. URL: https://npg.copernicus.org/articles/28/111/2021/

Ruckstuhl, Y.M. and Janjić, T. (2018) Parameter and state estimation with ensemble kalman filter based algorithms for convective-scale applications. *Quarterly Journal of the Royal Meteorological Society*, 144, 826–841. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3257

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*. URL: http://hdl.handle.net/hdl:21.14106/5c5f8a957ceea8e7d07e1ceb08cae837d90899a7.

Scheuerer, M. and Hamill, T.M. (2015) Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143, 4578–4596. URL: https://journals.ametsoc.org/view/journals/mwre/143/11/mwr-d-15-0061.1.xml

Selz, T. and Craig, G. (2021). *The transition from intrinsic to practical predictability of midlatitude weather*.

Stuart, A. and Ord, J. (2000) *Kendall's Advanced Theory of Statistics*. Arnold Publishers.

Williams, P.D. (2009) A proposed modification to the robert–asselin time filter. *Monthly Weather Review*, 137, 2538–2546. URL: https://journals.ametsoc.org/view/journals/mwre/137/8/2009mwr2724.1.xml

Williams, P.D. (2011) The raw filter: An improvement to the robert–asselin filter in semi-implicit integrations. *Monthly Weather Review*, 139, 1996–2007. URL: https://journals.ametsoc.org/view/journals/mwre/139/6/2010mwr3601.1.xml

Würsch, M. and Craig, G.C. (2014) A simple dynamical model of cumulus convection for data assimilation research. *Meteorologische Zeitschrift*, 23, 483–490. URL: https://doi.org/10.1127/0941-2948/2014/0492

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. URL: https://journals.ametsoc.org/view/journals/mwre/133/10/mwr3009.1.xml

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. (2002) The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, 83, 73–83. URL: http://www.jstor.org/stable/26215325

How to cite this article: Tempest, K.I., Craig, G.C. & Brehmer, J.R. (2023) Convergence of forecast distributions in a 100,000-member idealised convective-scale ensemble. *Quarterly Journal of the Royal Meteorological Society*, 149(752), 677–702. Available from: https://doi.org/10.1002/qj.4410

APPENDIX A

TABLE A1 Convergence measure data used for fitting of $y = an^{-1/2}$. Data up to a certain ensemble size cut-off (column 4) were not used in the fitting procedure

Model variable	Distribution	Statistic	Fitting cut-off
Wind	Figure 4c	Mean	0
		Variance	100
		0.6 quantile	0
		0.7 quantile	0
		0.95 quantile	100
		0.99 quantile	200
		0.999 quantile	2,000
	Figure 4h	Mean	0
		Variance	100
Height	Figure 5c	Mean	3
		Variance	100
		0.3 quantile	500
		0.375 quantile	80,000
		0.4 quantile	2,000
		0.6 quantile	30
		0.999 quantile	1,000
	Figure 5h	Mean	3
		Variance	100
Rain	Figure 6c	Mean	3
		Variance	100
		0.6 quantile	5
		0.7 quantile	5
		0.95 quantile	100
		0.99 quantile	300
		0.999 quantile	70,000
	Figure 6h	Mean	3
		Variance	100