RESEARCH ARTICLE

Journal Section

Convergence of forecast distributions in weak and strong forcing convective weather regimes

Kirsten I. Tempest | George C. Craig | Matjaž Puh | Christian Keil

Meteorological Institute, Ludwig-Maximilian-University Munich, Munich, Germany

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

Correspondence
Email: K.Tempest@physik.uni-muenchen.de

Funding information
The research leading to these results has been done within the subproject A6 of the Transregional Collaborative Research Center SFB / TRR 165 "Waves to Weather" (www.wavestoweather.de) funded by the German Research Foundation (DFG).

The constraint of computational power and the huge number of degrees of freedom of the atmosphere means a sampling uncertainty in probabilistic ensemble forecasts exists. In our previous study, uncertainty could be quantified by using a bootstrapping technique to create a Convergence Measure which converges proportional to $n^{-\frac{1}{2}}$ in the limit of large ensemble size. This power law can then be extrapolated to determine how sampling uncertainty would decrease with larger ensemble sizes and hence find the necessary ensemble size. It is unknown however how the sampling uncertainty depends on different weather regimes. This study extends the previous idealised ensemble developed, to include weak and strong forcing convective weather regimes to look at how sampling uncertainty convergence differs in each. Two $5,000$-member ensembles were run, with weak and strong forcing respectively. Comparisons with a kilometre-scale weather prediction model ensured realistic weak and strong forcing regimes by comparing the rain, CAPE, convective adjustment timescale and distribution shapes throughout the diurnal cycle. Differences in distribution shape between the regimes led to differences in the Convergence Measure. Large differences in spread between weak and strong forcing runs throughout the 24 hours led to large

differences in sampling uncertainty of the mean and standard deviation, which could be quantified according to well-known equations. The timing of these differences was case dependent. For extreme statistics such as the 0.95 quantile and for cases where there was precipitation, the moisture variables for the weak forcing case had the largest sampling uncertainty and required the most members to convergence proportional to $n^{-\frac{1}{2}}$. This was due to the tails of the weak forcing moisture variables containing the least amount of density. Different ensemble sizes will hence be required depending on whether one is in the weak or strong forcing convective weather regime.

KEYWORDS
idealised model, ICON, ensembles, weather prediction, asymptotic convergence, sampling uncertainty, distributions, weak forcing, strong forcing, convective weather regimes

1 | INTRODUCTION

Ensemble forecasting can only be precise to a certain extent due to the many degrees of freedom of the atmosphere and the limited computational power with which to predict it. This creates a considerable sampling uncertainty, especially with current operational ensemble sizes of between 20 and 50 members. As such, it is important to know how many members are necessary in an ensemble in order to reach the accuracy required as well as for a better distribution of resources to where they're most needed. Furthermore, it is suspected that different ensemble sizes may be required for different convective weather regimes that the atmosphere may be in.

Steps have been taken to investigate how sampling uncertainty may decrease as an ensemble acquires more members. Craig et al. (2022) used bootstrapping with replacement on their 1,000-member Numerical Weather Prediction (NWP) ensemble to estimate the sampling uncertainty of a specific statistical quantity and model variable. Bootstrapping created multiple statistically similar ensemble distributions so that the statistical quantity could be calculated multiple times, allowing for a 95% Confidence Interval (CI) to be constructed, which showed the sampling uncertainty associated with that statistic and variable. They discovered that for all distribution shapes and most forecast variables, the width of the CIs decreased proportional to $n^{-\frac{1}{2}}$ with increasing ensemble size n. This asymptotic scaling occurred as a result of those statistics tested obeying the Central Limit Theorem (CLT). The statistics that converged included the mean, which needed less than 10 members to converge, and the standard deviation, which needed between 50 and 100 members to converge. Convergence was not seen for the "extreme" 95$^{\text{th}}$ percentile however because of it's proximity to the tails of the distribution, meaning that more members would be required to resolve this. The model variables tested were the temperature at 2m and 500hPa, specific saturation deficit at 500hPa, as well as the probability of precipitation exceeding certain thresholds. These variables were chosen as they represented the three categories of distribution shape (quasi-Gaussian, multi-modal and highly skewed) which were found in the NWP forecast. If asymptotic convergence was reached, this would then allow for a framework to determine the ensemble

size needed to reach a specific level of sampling uncertainty by simply extrapolating the power law to smaller sampling uncertainties. This study was limited however as only a single case study was investigated. Furthermore, the $1,000$ member ensemble was not large enough to see convergence in statistics including the 95^{th} percentile. This would have required a larger ensemble.

The neighbourhood method has been shown to be successful in increasing the effective ensemble size of convective-scale ensembles (Craig et al., 2022; Puh et al., 2023b). It works by sampling grid points within a specified neighbourhood, rather than from a single grid point. The grid points are treated as individual ensemble members, increasing the effective ensemble size and providing additional information if the grid points within the neighbourhood are uncorrelated (Craig et al., 2022). By averaging out the uncorrelated small-scale noise among ensemble members in the neighbourhood, smoother distributions can then be created with reduced sampling uncertainty. To be effective, it is important that the ensemble members within the neighbourhood have similar statistical properties. Otherwise the distribution shape will change as the neighbourhood region becomes larger, incorporating members with different statistical properties. If a neighbourhood region is a circle, the statistical properties will often begin to become inhomogeneous at a radius of around 100km as a result of including different orographies and synoptic weather conditions. Although the neighbourhood method was employed on a 1000-member ensemble (Craig et al., 2022), a larger ensemble was still required to reach convergence proportional to $n^{-\frac{1}{2}}$ in extreme quantile statistics.

A $100,000$-member ensemble using an idealised model replicating convection (Tempest et al., 2023) was created to determine how sampling uncertainty converges with ensemble size for those variables not converged with the $1,000$-member ensemble (Craig et al., 2022). The idealised model was a modified single-layer shallow water model with three variables: wind, height and rain (Wuersch and Craig, 2014). Thresholds in the height, which once reached, altered the geopotential, allowed for updrafts and downdrafts of convective cells and as such for a model which replicated the basic processes of an NWP. In addition to the convergence proportional to $n^{-\frac{1}{2}}$ observed in Craig et al. (2022), in the idealised ensemble convergence was seen for extreme percentiles up to the 99^{th} percentile although tens of thousands of members were required for this. A major result of Tempest et al. (2023) was identifying a connection between the distribution shape of the model variables and the sampling uncertainty. Specifically, the equation for the sampling uncertainty of the quantiles (Walker, 1968):

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

was shown to produce estimates which were very close to the measured sampling uncertainty from the ensemble. Equation (1) scales as $n^{-\frac{1}{2}}$ and depends inversely on the density of the distribution f at q_p, the quantile of quantile level p. As such, it was concluded that the ensemble size required based on this framework depended on the model variable's distribution shape as well as the statistical quantity and the level of sampling uncertainty required.

A major factor affecting distribution shape is the convective weather regime. As distribution shape affects the convergence of sampling uncertainty, it is hypothesised that different convective weather regimes would have different convergence metrics and therefore require different sizes of ensemble to reach the same level of sampling uncertainty. Convective weather regimes relevant to the midlatitudes are weak and strong forcing and it is to our knowledge unclear how many ensemble members would be required for each. As well as Equation (1) which describes how asymptotic convergence of the quantiles depends on the underlying distribution shape, there are equations which describe how the sampling uncertainty (standard error) of the mean and standard deviation decrease with ensemble size. For the mean,

$$\sigma_m = \frac{s_X}{\sqrt{n}}, \tag{2}$$

where s_X is an estimate of the population's standard deviation. It can be seen that the sampling uncertainty decreases proportional to $n^{-\frac{1}{2}}$. A standard error exists for the standard deviation which also depends on the population's standard deviation and decreases proportional to $n^{-\frac{1}{2}}$:

$$\sigma_{sd} = \sqrt{\frac{\pi}{2}} \frac{s_X}{\sqrt{n}}, \tag{3}$$

although this requires normality of the underlying distribution. As such it could be expected that if the spread and density distributions of the model variables between the convective regimes vary, so will the convergence behaviour and the size of ensemble required.

Weak and strong forcing regimes can often be categorised depending on how convection is initiated and maintained. Convection in a weak forcing regime occurs when large-scale processes have built up CAPE over a timescale which is long compared with the timescale with which the instability is removed because of significant Convective Inhibition (CIN) in the atmosphere. This CIN could be for example an inversion, where warm, light air is above cold, dense air (Keil et al., 2014). A "trigger" that could be for example, orography or solar insolation, eventually overcomes this inhibition and allows for the release of CAPE by latent heat release which is often in the form of many precipitating single convective cells. As a result of solar insolation playing a role in weak forcing scenarios in triggering the release of CAPE, there is often a characteristic diurnal cycle with significant precipitation around midday and less in the night. In the strong forcing regime, there is little CIN and so when CAPE is produced by large scale processes it is quickly consumed by latent heat release. This is known as equilibrium convection and can be maintained much longer than convection in the weak forcing scenario. Moist convection in these strong forcing regimes usually has the form of larger convective structures such as squall lines.

One method to differentiate between convective regimes is by analysing the convective adjustment timescale (Done et al., 2006). The timescale is an estimate of how quickly convection consumes CAPE and is the ratio of the convective instability (CAPE) to the rate of its removal by the convection (stabilisation) (Keil et al., 2019):

$$\tau_c = \frac{\mathrm{CAPE}}{\frac{d\mathrm{CAPE}}{dt}}. \tag{4}$$

If there is CIN in the environment, as in a weak forcing scenario, CAPE will be able to increase to large values until something triggers its release. Furthermore, the eventual latent heat release in the weak forcing scenario is stronger due to the larger amounts of CAPE allowed to be built up, providing a large denominator value. Overall however, there are larger values for τ_c in weak forcing scenarios than if there is less CIN and the CAPE is continually removed by convection, meaning that it cannot build up to such high values, as in the strong forcing scenario. A threshold (usually between 3 and 12 hours (Zimmer et al., 2011)) is set in a model to distinguish between the weakly and strongly forced convective weather regimes.

The two components of the convective adjustment timescale (τ_c), the CAPE and precipitation, show characteristic behaviours in weak and strong forcing regimes. Examples of the weak and strong forcing regimes are shown in Figure 1,

using data from 20-member ensembles from ICON-D2, a full convective-scale NWP forecast (Puh et al., 2023a). It shows the weak forcing day of 10 June and the strong forcing day of 29 June over Germany. Beginning at 00:00 UTC in the weak forcing, it can be seen that the CAPE is relatively small, as well as the rain. Then as solar insolation increases in the morning, the CAPE increases and due to turbulence in the boundary layer from the thermal energy, convection begins soon thereafter. As convection reduces the buoyancy in the domain, the CAPE decreases slightly, as seen around 12:00 UTC. Due to the solar insolation however, the CAPE only decreases strongly after 18:00 UTC. The rain has reached its peak at around 12:00 UTC and decreases thereafter. Apart from the dip at 04:00 UTC, the convective timescale is very large (approximately 60 hours) until convection begins, but then decreases to approximately 10 hours thereafter. In strong forcing on the other hand, the convective timescale is relatively low, under 4 hours, throughout the 24 hour forecast. Although the CAPE is larger in general for the strong forcing and more constant, the diurnal cycle is still seen as it obtains higher values during the middle of the day. The precipitation also has higher values as well as being more constant throughout the day, although it is seen to dip in the morning. This is due to a cold weather front passing through Southern Germany in the afternoon. Although these two scenarios will not be precisely replicated, these examples show the broad characteristics of weak and strong forcing and will be expected to be replicated in the model used in this study.

Due to the interaction of the larger-scale flow with convection, strong forcing is generally more predictable than weak forcing i.e. the location and intensity of the convection can be predicted more accurately for a longer period of time. The predictability is often measured by the spread of the distribution of the forecast ensemble of a certain variable, for example precipitation. Keil et al. (2014) analysed 88 days during the Summer of 2009 using a NWP ensemble which has a grid size small enough to broadly resolve convective motions and found that the predictability was higher for hourly total precipitation when the strong forcing regime was dominant. Bachmann et al. (2020) implemented two other predictability measures, the believable and the decorrelation scale which measures at what small scale the ensemble still represents the observations and at what scale the ensemble members become decorrelated, respectively. Data was analysed from three summer periods simulated using an operational ensemble forecast and it was likewise found that strong forcing regimes had greater predictability than weak forcing regimes.

The differences in predictability between the weak and strong forcing regimes can be attributed to the spread of the underlying model variable distributions, indicating that the distribution shapes depend on the forcing regime. As the nature of the convergence of sampling uncertainty with ensemble size is known to depend on the underlying ensemble distribution shape, it is plausible that one forcing situation may be more prone to larger sampling uncertainty than the other. Specifically what we might expect is that in periods of strong precipitation, the tails of the weak forcing moisture (height and rain) variable distributions are longer and less dense, leading to a larger sampling uncertainty for extreme quantiles following Equation (1). As well as a larger sampling uncertainty in this scenario, it could be expected that the weak forcing would need more ensemble members for the convergence of sampling uncertainty to scale proportional to $n^{-\frac{1}{2}}$, if the distribution is less defined at extreme values. Moreover, in the early hours and in the evening, strong forcing would be expected to have a larger spread in its variables, which would mean a larger sampling uncertainty in the mean according to Equation (2). This would also apply to the standard deviation statistic described by Equation (3), however only in cases where the underlying distribution is normally distributed. The question is then how does convergence of sampling uncertainty with ensemble size differ between the convective weather regimes of weak and strong forcing throughout a diurnal cycle, and what does this mean for the ensemble size required?

In this paper, we investigate the difference in convergence of sampling uncertainty with ensemble size between weak and strong forcing using a similar approach to Tempest et al. (2023). That is, an idealised model will be used to create large ensemble sizes. After its success in Craig et al. (2022), the neighbourhood method will be used in combination with the large ensemble size to increase the effective ensemble size further. The same idealised model

will be used as in Tempest et al. (2023), however it will be extended to allow for different convective weather regimes. As such, the requirements for the model are that it has space and time scales representative of convective processes, can model non-linear processes, be computationally efficient as well as be able to accurately portray the differences between weak and strong forcing as highlighted in Figure 1. Furthermore, the spread in the weak forcing should be larger than the strong forcing amongst moisture model variables at times when significant precipitation occurs. Using the extended idealised ensemble, two experiments will then be run, one with weak forcing and the other with strong forcing. In order to ensure the results from the idealised model are realistic enough, the distributions from weak and strong forcing runs from ICON-D2 will be analysed and compared to the idealised model results. Differences between the distributions in the weak and strong regimes will be highlighted as areas of potential difference in the convergence of sampling uncertainty. A Convergence Measure (Tempest et al., 2023) will then be used to identify differences in the convergence of sampling uncertainty between the two convective weather regimes and it will be verified whether these relate to the differences seen in the distribution shapes.

The paper is organised as follows. Section 2 will introduce the extended idealised ensemble which replicates weak and strong forcing along with the methods used to analyse the ensemble data. The distributions will then be analysed in Section 3 to ensure the model is replicating realistic behaviour and because it is expected that distribution shape is a key factor in explaining the differences in sampling uncertainty between the forcing regimes. Then the Convergence Measure will be used in Section 4 to ascertain whether the different forcing regimes do lead to different convergence behaviours and ensemble sizes required. Discussion and conclusions will then follow in Section 5.

2 | MODEL AND METHODS

2.1 | Extended idealised model

The idealised model from Tempest et al. (2023) is extended to allow for weak and strong forcing regimes. Figure 1 which uses data from a full NWP model will be used as a reference in this adaptation. Unless stated, the idealised model remains as in Tempest et al. (2023). As it is only extended, it is expected that the model will satisfy the conditions of having space and time scales representative of convective processes, can model non-linear processes and be computationally efficient.

It is known that the rate at which CAPE is consumed in the atmosphere by convection can differ between weak and strong forcing regimes. As such, a measure of CAPE is introduced into the idealised model as the first step in extending it. CAPE effectively measures the buoyancy of air, therefore the equivalent of CAPE in the idealised model is the difference between the geopotential (ϕ) and the constant geopotential (ϕ_c), which a grid point acquires if it is a cloud (above the H_c threshold) and which then allows for buoyancy and the developing, updraft phase of a cloud. It can be written as:

$$\text{CAPE}_e = (g\bar{h} - \bar{\phi_c})_e, \tag{5}$$

where the bars are domain averages and $_e$ indicates an ensemble average. The CAPE does not remain constant however, and so a time dependant ϕ_c is required. ϕ_c will decrease when there is more solar insolation or stronger synoptic forcing, allowing for more buoyancy, and increase with any convection, creating less buoyancy in the model. How CAPE evolves can then be used to create either a weak or strong forcing regime in the model. The time-dependent equation for ϕ_c is:

$$\frac{d\phi_c}{dt} = S_{\text{rad}} - S_{\text{for}} + S_{\text{con}}, \tag{6}$$

where the terms on the right hand side control the solar insolation to allow for a diurnal cycle (S_{rad}) and dictate the strength of synoptic forcing (S_{for}) as well as convection (S_{con}).

Solar insolation begins during the morning and increases until 12:00 UTC and then decreases thereafter. As such, a time dependence is required in the form of a cosine and the condition that there is no solar insolation before 06:00 UTC and none after 18:00 UTC. Therefore:

$$S_{\text{rad}} = \left\{ \begin{array}{cc} a_r \cos\left(\frac{2\pi t}{t_{\text{total}}}\right), & \frac{t_{\text{total}}}{4} < t < \frac{3t_{\text{total}}}{4} \\ 0, & \text{otherwise} \end{array} \right\}, \tag{7}$$

where a_r has the value $2.5 \cdot 10^5 \text{m}^2\text{s}^{-3}$. t is the current time step and t_{total} are the total number of time steps in one diurnal cycle.

The forcing term is likewise also time dependent to mimic the front passing through Southern Germany in the afternoon of 29 June 2021. It is tuned so that in the case of strong forcing, a front lasts for approximately 13 hours, with the peak occurring at 18:30 UTC:

$$A_{\text{for}} = a_f \left(1 + \sin\left(\frac{2\pi(t - T_{\text{shift}})}{T_{\text{for}}}\right)\right) \tag{8}$$

$$S_{\text{for}} = \left\{ \begin{array}{cc} A_{\text{for}}, & A_{\text{for}} > a_f \\ a_f, & \text{otherwise} \end{array} \right\}. \tag{9}$$

where each ensemble member has a random a_f, chosen from a Gaussian distribution of mean $1.5 \cdot 10^{-5}$ and standard deviation of $1.5 \cdot 10^{-6}$ in the case of strong forcing. T_{for} is set to a period of 26 hours and T_{shift} provides a shift of 11 hours.

The convection term produces negative buoyancy in the case of rain and convergence in the wind. It depends on the strength of convergence and is tuned with γ_2 so as to balance it with the buoyancy terms of S_{rad} and S_{for}.

$$S_{\text{con}} = \left\{ \begin{array}{cc} \gamma_2 \beta \frac{\bar{du}}{dx}, & Z > H_r \text{ and } \frac{du}{dx} < 0 \\ 0, & \text{otherwise} \end{array} \right\} \tag{10}$$

where γ_2 is set to $-2000 \text{m}^2\text{s}^{-2}$ and β is 0.1 as before. To disincentivise clouds to form at grid points which recently had a cloud, there will be cases where there is no convection but rain still in the domain. This mimics the effects of cold pools (Hirt et al., 2020).

To incorporate the diurnal cycle, whereby solar insolation encourages convection during the day, the stochastic forcing in the original model which acts on the height field at every time step at a random grid point is made to be

time dependent. The magnitude of the perturbations is multiplied by a constant, C_{stoc}, which is:

$$C_{\text{stoc}} = \left\{ \begin{array}{cc} -\cos\left(\frac{2\pi t}{t_{\text{total}}}\right), & \frac{t_{\text{total}}}{4} < t < \frac{3t_{\text{total}}}{4} \\ 0, & \text{otherwise} \end{array} \right\}. \tag{11}$$

The addition of Equation (6) hence allows for ϕ_c to change in time, allowing for periods of greater buoyancy and periods with less buoyancy. The stabilising term (S_{con}) increases ϕ_c, making the model more stable whereas the destabilising terms of S_{rad} and $-S_{\text{for}}$ decrease ϕ_c, making the model less stable and encouraging more convection to reduce the instability. When CAPE is large and there is not a lot of convection, the model is not in equilibrium, and this would be a weak forcing scenario. Only when there is a trigger, such as orography or solar insolation, could the CAPE be released in the form of convection. When the atmosphere is in equilibrium, convection removes CAPE when it is created. This would be a strong forcing regime.

As explained previously, the convective timescale (Equation (4)) can also be used to differentiate between weak and strong forcing regimes. It is the speed at which convection removes CAPE and can be calculated in the extended idealised ensemble by:

$$\tau_c = \frac{\text{CAPE}_e}{3600 \cdot S_{\text{con}_e}}, \tag{12}$$

where the division by 3600 seconds means that the units of τ_c is hours.

2.2 | Set-up of extended idealised ensembles

Two ensembles were created with the extended idealised model, namely a weak forcing run and a strong forcing run. Any differences to how the idealised 100,000-member ensemble was implemented by Tempest et al. (2023), are detailed here. For the initialisation, the domain for the two experiments was at rest initially with zero rain and wind and the height set at 38m, for all grid points. The fluid depth was lowered as the gravity wave speed was too fast and creating too much convection at later time points in the weak forcing case when the fluid depth was at the previous level of 90m. For both experiments, a 1,000-member ensemble was used as the input background to the ensemble Kalman filter (EnKF) Data Assimilation (DA) (Evensen, 1994) which was said to be at 00:00 UTC. The DA then used the respective model for either weak or strong forcing and was cycled 288 times with 75 time steps between each cycling. This covered a time period of 24 hours and so captured one diurnal cycle. The set-up of the DA was the same as for the 100,000-member ensemble however with the adaptations to the model as mentioned above. The free-runs for the weak and strong forcing ensembles were then run for 24 hours, being initialised using the analysis from the DA. The ensembles were then extended to a size of 5,000 each for the free-run through copying the analysis ensemble members 5 times.

In the set-up of the extended idealised ensembles, certain model parameters were re-tuned due to the addition of Equation (6) and the lowering of the fluid depth. This included the amplitude of the random perturbations to be increased to 0.011 from 0.00895. The thresholds for clouds and rain were decreased to 38.02m and 38.4m respectively and the wind (K_u) and height (K_h) diffusion constants were changed to $3.4 \cdot 10^3 \text{m}^2\text{s}^{-1}$ and $1.4 \cdot 10^3 \text{m}^2\text{s}^{-1}$ respectively. Furthermore, a cut-off for the rain variable was introduced. When the rain was below 0.000009, it was automatically decreased to zero. This was to improve how realistic the idealised model was as rain below that threshold was

negligible and did not impact the operation of the model.

2.3 | Properties of the extended idealised ensemble

The idealised ensemble has now been extended to allow for weak and strong forcing regimes. Here it is checked that this is done accurately, by analysing the evolution of CAPE, rain and the convective timescale throughout one diurnal cycle for the weak and strong forcing ensemble runs and comparing these with the ICON-D2 model output (Figure 1).

The evolution in time of CAPE, the convective timescale, as well as the total rain in the domain of the extended idealised ensemble for the two forcing regimes can be seen in Figure 2. In the weak forcing case, the CAPE is low after 00:00 UTC and begins to increase after 06:00 UTC when the solar insolation term begins to increase, making the atmosphere less stable. Similarly, there is basically no rain in this time period. As such, the convective timescale is very high. After about 09:00 UTC there is enough instability for the first rain clouds to be created. Until about 12:00 UTC, the CAPE has continued to increase but after this point the solar insolation begins to decrease and the convection is also very strong at this point, both acting to reduce the CAPE. The rain continues to be dominant until approximately 15:00 UTC where it has then exhausted the CAPE in the atmosphere and begins to decrease. The CAPE and rain then slowly decrease from their peaks. The convective timescale has reduced to a small value of approximately 10 hours after the beginning of the convection and remains around this value for a period of time. As the rain begins to wane, the convective timescale gradually begins to increase again, however not to the previously high values. In the strong forcing scenario, the CAPE is relatively constant throughout the 24 hours and at all points there is a large amount of rain. This makes for a relatively constant low value for the convective timescale, as expected. In this strong forcing experiment one still sees the diurnal cycle in the rain.

The evolution of the three quantities of the CAPE, convective timescale and precipitation in Figure 2 for the weak and strong forcing runs are in line with those computed from the ICON-D2 runs for 10 and 29 June 2021 respectively (Figure 1). In both the idealised and NWP ensemble, the diurnal cycle is clearly seen in the weak forcing run and to a lesser extent in the strong forcing. As well as this, the general time evolutions of the three quantities match.

The spread in the ensemble is additionally important for differentiating between weak and strong forcing. The spread of the rain and CAPE values appears consistently larger in the idealised strong forcing case of Figure 2. However, if one measures the standard deviation (spread) across the extended idealised ensemble for every individual grid point and then averages across the domain, the weak forcing has a larger spread in the moisture variables for a 5 hour period around 12:00 UTC, as would be expected due to the sporadic nature of the weakly forced precipitation.

Through analysing and comparing Figures 1 and 2 and checking the spread, the extended idealised ensemble has been shown to produce sufficiently realistic weak and strong forcing regimes. Furthermore, it has been deemed to be realistic enough in terms of the space and time scales being representative of convective processes, can model non-linear processes and be computationally efficient, from carrying out similar analysis as in Tempest et al. (2023) (not shown). Still of interest to analyse to ensure the accuracy of this extended idealised ensemble and the variation between the weak and strong forcing regimes, is the evolution of the model variable distributions from throughout the 24 hour free run.

2.4 | Statistical analysis

2.4.1 | Neighbourhood method

To increase the effective ensemble size in this study, the neighbourhood method is employed. The neighbourhoods were chosen so to be centred close to the middle of the domain, at grid point 501 of 1,000 (beginning at 1). Three neighbourhood sizes were then created, by including x grid points above and below grid point 501 on the one-dimensional domain. x was 2, 10 and 20. This gave neighbourhood sizes of length 2km, 10km and 20km respectively.

2.4.2 | Convergence Measure

The Convergence Measure developed in Tempest et al. (2023) is used to estimate how sampling uncertainty decreases with ensemble size. It is the width of the 95% CI of the sampling distribution of a chosen statistic created from 10,000 bootstraps (with replacement) of the original ensemble distribution of the chosen variable. To show the power-law behaviour of asymptotic convergence, it is plotted on a log-log scaled axis with ensemble size on the x-axis.

3 | DISTRIBUTIONS FROM WEAK AND STRONG FORCING RUNS

The distribution shapes are now analysed for the weak and strong forcing runs throughout their 24 hour free run forecast to ensure they are realistic enough for our purposes of creating forecast runs which differentiate between the two regimes. The differences in distribution shape are furthermore expected to be important in explaining differences in the Convergence Measure. Expectations from the analysis of the distributions will therefore later be compared with the Convergence Measure calculated from the distributions.

Several characteristics of the distribution shapes are expected. First, it is expected that the three categories of distribution shape which are common to convective-scale weather forecasting (quasi-Gaussian, multi-modal and highly skewed) (Craig et al., 2022), are seen. Furthermore, for the forecast to realistically portray weak and strong forcing, it is anticipated that the weak forcing runs begin with little spread in all three model variables. At 06:00 UTC and onwards significant spread would be obtained before decreasing again in the afternoon. This strong diurnal cycle behaviour is not expected to be as obvious in the strong forcing. Furthermore, longer and thinner tails are anticipated in the weak forcing for the moisture variables during periods of heavy precipitation.

To show the time evolution of the distribution shapes, Figure 3 shows contour histogram plots of the 24 hour evolution of the distributions using a neighbourhood of length 20km of the three model variables (columns) from the weak and strong forcing runs (rows). The general shape of the distributions were similar for all neighbourhood sizes and the single grid point, however the larger the neighbourhood, the smoother the distributions became (not shown). First of all it can be confirmed that at all times the three expected distribution shape categories of Gaussian, multi-modal and highly skewed were produced by the ensembles. Examples of Gaussian distributions can be seen in the wind at all time points and forcing types. Bi-modality can be seen in the height, for example after 12:00 UTC in the weak forcing. Highly skewed shapes can then be seen in the rain, for example after 18:00 UTC in the weak forcing scenario. This indicates that similar, realistic, distributions as in Tempest et al. (2023) are produced by the extended version of the idealised model.

Employing Figure 3, the spread of the forecast distributions is now analysed. The diurnal cycle is seen from the changes in spread in all model variables of the weak forcing and is not as prominent in the strong forcing, as expected. The height variable however shows no decrease in spread as the wind and rain variables in the evening. The spread

is calculated in Table 1 for the time points of 06:00, 12:00 and 20:00 UTC, as these time points contain all types of distribution shapes seen in the diurnal cycle. Note that the same distribution shapes are seen at 12:00 UTC as at 15:00 UTC when rain in the domain reaches its maximum. It is clear from Table 1 that in the weak forcing for all three model variables, the spread is very small compared to the strong forcing at 06:00 UTC and slightly smaller at 12:00 UTC and at 20:00 UTC. After 18:00 UTC the spread decreases for the wind and rain variables however not for the height in the weak forcing case. The spread becomes similar in the evening for both forcing runs. The spread at 12:00 UTC was larger in the strong forcing than the weak forcing for the moisture variables, not as expected. When normalised however, the rain has a larger spread in the weak forcing at 12:00 UTC. Nevertheless, the expected behaviour of the spread being very small for the weak forcing compared to the strong forcing in the morning is seen, as well as the evolution of the weak forcing increasing in spread quickly as rain occurs around 12:00 UTC.

The tails of the distributions from the weak and strong forcing runs are now compared in Figure 3. The probability density of the 0.95 quantiles are listed in Table 2 and are visualised by black dots for each variable and for both forcing runs. It is seen that the density of the 0.95 quantile for the wind is relatively similar between weak and strong forcing but that there are larger differences for the moisture variables. The density of the 0.95 quantile is smaller for the weak forcing case for the moisture variables when there is a lot of precipitation at 12:00 UTC. This is also the case for the height at 20:00 UTC when there are still a significant number of clouds in the domain. This indicates that in general, during time periods of significant precipitation, the tails of distributions of moisture variables in the weak forcing regime are longer and less dense than those in the strong forcing regime.

To test how realistic the idealised version is in creating weak and strong forcing regimes, distributions from the extended idealised ensemble are compared with those from the 120-member ICON-D2 ensemble. The height and rain variables from the idealised ensemble correspond to those of the relative humidity and precipitation variables from ICON-D2 respectively. Low values of the relative humidity variable indicate relatively dry air, and higher variables indicate wetter air, with saturation occurring at 100%. In Figure 4 it is seen that for both variables, the spread increases significantly in the first 14 hours for both forcing regimes. Although more time is needed in the weak forcing until the maximum spread of the relative humidity is reached. The spread of the relative humidity for the weak and strong forcing regime is most divergent in the mid morning around 06:00 UTC, and then becomes closer together as the day goes on, as seen in the idealised ensemble. For the total precipitation however, the differences in spread between the weak and strong forcing become larger throughout the day and then decrease towards the evening again. The time evolutions in spread seen from the idealised and full numerical model ensembles suggests that the differences between weak and strong forcing depend largely on the specific cases selected. For example if the cold front passed through in the morning of 29 June, large amounts of rain in the early morning would lead to the spread having large differences between weak and strong forcing at this time, and then would get more similar throughout the day. This is seen in the idealised ensemble, where there is more constant rain than in the ICON-D2 simulation for the strong forcing case. Nevertheless, in both the idealised and ICON-D2 ensemble, there have been seen to be large difference in spread in weak and strong forcing runs and this will likely have consequences for the Convergence Measure.

Comparing the tails of the moisture variable distributions between the idealised and ICON-D2 ensemble, similarities are observed. Using Figures 3 and 4, it can be seen that in both ensembles the tails for the weak forcing moisture distributions during time periods of precipitation are longer and less dense, although of a smaller magnitude, compared to the strong forcing. The time point of 14:00 UTC is used for comparison as a time when large amounts of precipitation was occurring in both the weak and strong forcing cases. From comparing the probability densities between weak and strong forcing at the 0.95 quantile for ICON-D2 (black dots in Figure 4), it has been confirmed that even during periods of strong precipitation in the strong forcing, the weak forcing remains to have the smallest density, and therefore the thinnest distribution tails. Although the distributions are not exactly comparable between the idealised

and ICON-D2 ensemble, important similarities of the spread and the density of the tails exist which are expected to lead to distinct behaviour in the Convergence Measure for weak and strong forcing. As such, it can be concluded that findings from the idealised model which compare weak and strong forcing can be applicable in a broader context to larger, more complex NWP models. It has in general been seen that the time evolution and shape of the distributions from the extended idealised ensemble are as to be expected and that they furthermore show differences between weak and strong forcing which are also seen in the ICON-D2 ensemble distributions. That is, the distributions from the extended idealised ensemble convey important differences between weak and strong forcing which will likely be significant in creating different characteristics in the convergence of sampling uncertainty with ensemble size since the Convergence Measure is dependent on the underlying distribution shape. In particular, the spread is seen to be different at many time points between the weak and strong forcing cases. For the case of the idealised ensemble, the spread is very small for the weak forcing in the morning and then increases quickly as convection begins, whereas the strong forcing does not show as much diurnal variability. It is therefore expected that, according to Equation (2), that the Convergence Measure for the mean of the weak and strong forcing in the morning will have very different values, with strong forcing having the largest uncertainty for this time period. This difference in sampling uncertainty would then decrease as convection begins, but as strong forcing has the largest spread, it will consistently have the largest sampling uncertainty for the statistic of the mean. As the distribution shape has a similar influence on the standard deviation as for the mean according to Equation (3), similar convergence behaviour is expected in this case, but only for the wind variable as the condition of Gaussianity is required. The second main difference in the distributions between weak and strong forcing is that in periods of heavy precipitation, the tails of the moisture variables in the weak forcing are longer and therefore less dense at extreme values. Following Equation (1) this will lead to larger sampling uncertainty in the weak forcing during these convective periods. In addition it would be expected that more ensemble members would be required for the Convergence Measure to converge proportional to $n^{-\frac{1}{2}}$. In the following, it will be analysed as to whether these differences in the distributions are observed in the Convergence Measure.

4 | SAMPLING UNCERTAINTY CONVERGENCE

Due to the different distribution shapes seen in weak and strong forcing regimes throughout one diurnal cycle, it is expected that there will be different sampling uncertainty behaviour for each regime. The magnitude of the sampling uncertainty as well as how quickly it converges proportional to $n^{-\frac{1}{2}}$ will be investigated. For this, the Convergence Measure will be analysed and particular attention given to the variables and statistics where particular differences between the weak and strong forcing distributions were observed. The statistical quantities are analysed based on knowledge of how certain aspects of the distribution shape can affect that statistic e.g. the standard deviation of the distribution can impact the sampling uncertainty of the mean according to Equation (2).

Figure 5 shows the Convergence Measure for the mean statistic for the three model variables and three time points using data from a neighbourhood of length 20km. These three time points were chosen because they each had different distribution shapes and were at different phases of the diurnal cycle. As expected for the mean (Tempest et al., 2023), the Convergence Measure is scaling proportional to $n^{-\frac{1}{2}}$ with less than 10 ensemble members. At 06:00 UTC, the Convergence Measures for the weak and strong forcing are quite different for the wind and height (there is no rain at this time in weak forcing). As convection begins, these lines get closer together and this continues into the evening. This is expected according to Equation (2), whereby the larger the standard deviation of the underlying distribution, the larger the uncertainty of the mean. As the standard deviations of the underlying distributions from the weak and strong forcing (Table 1) become closer together throughout the day, so does the sampling uncertainty

of the mean.

In a similar manner as the mean, the standard deviation Convergence Measure for the three model variables is shown in Figure 6. The same behaviour as for the mean is seen in the standard deviation for the wind whereby the Convergence Measure becomes closer and closer during the day. This is also due to the standard deviation of the underlying distribution becoming closer together, as Equation (3) states that the smaller the spread of the underlying distribution, the smaller the sampling uncertainty of the standard deviation. This is however not the case for the height and rain, where it is seen at 20:00 UTC that their weak and strong forcing Convergence Measures become more offset despite their spread becoming more similar. This is due to the height and rain distributions not satisfying the condition of normality required for Equation (3).

The Convergence Measures for the 0.95 quantile are shown in Figure 7 for the three model variables and three time points using data from a neighbourhood of length 20km. As a consequence of the density of the 0.95 quantile of the wind being consistently smallest for the strong forcing compared to the weak forcing as shown by Table 2, the strong forcing has the largest sampling uncertainty at all time points. Due to Equation (1), the smaller the density, the larger the sampling uncertainty of that specific quantile. The strong forcing does not always have the largest sampling uncertainty however. Concentrating on the moisture variables, the magnitude of the sampling uncertainty of the height variable during weak forcing is larger than the strong forcing at 12:00 and 20:00 UTC, and also in the case of the rain variable at 12:00 UTC. This was expected as for these variables at these time points, as the probability density at the 0.95 quantile is smallest for the weak forcing (Table 2) due to the long tails in the distribution. The sampling uncertainty is likely only largest for the weak forcing during these time points as significant precipitation is occurring then. In the weak forcing case, until late morning there are not many rain clouds and then suddenly many. This means that there will be a large density of noncloudy members in the ensemble, while other members acquire lots of rain. This creates a tail with less density than for the strong forcing, where more ensemble members would have already acquired clouds and rain earlier in the forecast. At 20:00 UTC the rain has decreased significantly in the weak forcing case, meaning that the tail of the distribution is no longer less dense than that in the strong forcing case. This is not the case for the height, where many clouds still exist in the weak forcing case, meaning that at 20:00 UTC the weak forcing still has a larger sampling uncertainty for the 0.95 quantile of height but not for the rain.

In contrast to results so far that considered neighbourhoods of length 20km, the Convergence Measure from a single grid point is now analysed in order to determine how distribution shapes affect how many ensemble members are required for convergence proportional to $n^{-\frac{1}{2}}$. Figure 8 shows the Convergence Measure for the 0.95 quantile as in Figure 7 but for a single grid point rather than for a neighbourhood. At 06:00 UTC, all measures are converging proportional to $n^{-\frac{1}{2}}$, as well as all wind variables with an ensemble containing $5,000$ members. It is seen however for distributions with long, low density tails, that they need significantly more members for their sampling uncertainty to converge proportional to $n^{-\frac{1}{2}}$. This is evident for the height variable in the weak forcing regime as it has not converged asymptotically at 12:00 and 20:00 UTC with $5,000$ members, but the strong forcing has. This corresponds to periods when the tail of the respective weak forcing distributions were the longest and least dense. Similarly for the rain: when the tail of the underlying distribution was very long, more members were needed for convergence to be proportional to $n^{-\frac{1}{2}}$. This is seen at 12:00 UTC during the weak forcing where the rain distribution's tail was less dense than the strong forcing's, leading to the weak forcing needing more than a $1,000$ members to converge asymptotically and the strong forcing significantly less. The opposite is true at 20:00 UTC when there was no longer significant rain in the weak forcing. This example demonstrates how distributions with longer and less dense tails can lead to needing more members until asymptotic convergence of sampling uncertainty proportional to $n^{-\frac{1}{2}}$ is observed.

Through analysing specific cases of the Convergence Measure and their underlying distributions, differences in the convergence of sampling uncertainty in weak and strong forcing regimes have been found. It has been seen

that in weak forcing cases with precipitation, the moisture variables will have a greater sampling uncertainty for their extreme quantile statistics and that they will need more ensemble members for convergence proportional to $n^{-\frac{1}{2}}$ to be observed compared to strong forcing regimes. Furthermore, the difference in sampling uncertainty between weak and strong forcing for the mean (and standard deviation in case of a Gaussian underlying distribution) has been found to be significant at particular times of the day when the distributions from the weak and strong forcing runs show the most difference in spread. For the idealised ensemble, the difference in spread between the weak and strong forcing across all model variables was greatest in the early morning, giving the strong forcing a much larger sampling uncertainty than the weak forcing at this time, and this difference then decreased throughout the day as the weak and strong forcing distributions became more similar. The convergence behaviour observed here was expected from the characteristics of the model variable distributions in combination with the equations for the sampling uncertainty of the mean, standard deviation and quantiles. As the distribution characteristics which led to these conclusions were also seen in data from ICON-D2, it is expected that these convergence results will additionally hold for full NWP ensembles.

5 | CONCLUSIONS

The question explored in this study was whether the ensemble size required to satisfactorily predict the weather would vary depending on if there was a weak or strong forcing convective weather regime. To investigate this, the convergence of sampling uncertainty with ensemble size in the weak and strong forcing regimes were compared.

Large idealised ensemble experiments with weak and strong forcing were simulated to explore how sampling uncertainty converges in both regimes. This involved extending the idealised ensemble from Tempest et al. (2023) by adding in an extra time dependent equation for the constant geopotential term which was made to depend on the solar insolation, synoptic forcing and convection. The solar insolation and synoptic forcing increased, while the convection damped the buoyancy a given cloud would have in the model. Equivalent terms for the CAPE and convective adjustment timescale were created in order to categorise the convective forcing regime within which the model was in. From comparisons of these two measures as well as the time evolution of the precipitation over a diurnal cycle with weak and strong forcing forecasts from the full convective-scale NWP ensemble, ICON-D2, it was ascertained that distinct forcing regimes had been created by the extended idealised ensemble. Following Tempest et al. (2023), the extended version of the ensemble continued to have space and time scales representative of convective processes, be able to model non-linear processes as well as be computationally efficient. A weak and strong forcing ensemble forecast was then created of length 24 hours beginning at 00:00 UTC with initial conditions provided by DA. Each ensemble had $5,000$ members in the free run.

The distributions were first checked to ensure they were sufficiently realistic as well as to observe how they varied between weak and strong forcing regimes. For this, the neighbourhood method was employed to expand the ensemble further. Three categories of distribution shape, quasi-Gaussian, multi-model and highly skewed were seen, in line with previous convective-scale studies (Craig et al., 2022; Kawabata and Ueno, 2020). Furthermore, behaviour unique to weak and strong forcing seen in the extended idealised ensemble were also seen to occur in ICON-D2. This indicated that the extended idealised ensemble was realistic enough for the purposes of differentiating between the forcing regimes. The differences seen in the different regimes amongst the distributions included large variations in spread between the weak and strong forcing distributions and longer, less dense tails in the weak forcing moisture distributions compared to the strong forcing during periods of precipitation.

The Convergence Measure was used to show differences in the nature of the convergence of sampling uncertainty

between the weak and strong forcing regimes. Specific differences arose, which were clearly linked with the underlying distribution shape. The mean statistic, which has larger sampling uncertainty when the underlying distribution has more spread according to Equation (2), had consistently larger sampling uncertainty for the strong forcing case since it's spread was larger. The difference in sampling uncertainty of the mean between the weak and strong forcing varied significantly, and this was related to the spread of the underlying distributions. Large variations in sampling uncertainty were also seen for the standard deviation, however this behaviour could not be linked well to the distribution shapes for the height and rain variables as normality was required for Equation (3) to apply. Longer, thinner tails in the weak forcing moisture distributions during time periods of precipitation led to larger sampling uncertainty in the extreme quantiles of the weak forcing compared to the strong forcing following Equation (1). Furthermore, it was seen that the longer the tail, the more members were needed to resolve the distribution shape and as such the more members that would be needed for convergence proportional to $n^{-\frac{1}{2}}$ to be observed.

The highlighted differences of how sampling uncertainty converges with ensemble size between weak and strong forcing regimes indicate that different ensemble sizes may be required depending on whether one is in the weak or strong forcing convective weather regime. If one is following the framework to determine ensemble size which was developed in Tempest et al. (2023) and Craig et al. (2022), one can find the desired ensemble size for each regime by calculating the Convergence Measure. Say our forecaster with a 50-member ensemble is interested in the accuracy of the 95^{th} percentile of rain over Munich and they want to know how many members are required to reach a certain level of sampling uncertainty for days when there is weak forcing and for days when there is strong forcing. They would begin by calculating the Convergence Measure for the 95^{th} percentile of the rain. The neighbourhood method or the parameterisation method (Tempest et al., 2023) would likely be needed for a large enough effective ensemble size for it to be converging proportional to $n^{-\frac{1}{2}}$. They would then extrapolate this Convergence Measure to smaller values of sampling uncertainty until they were satisfied and the corresponding ensemble size would be their required size of ensemble for that level of sampling uncertainty. To measure the ensemble size required in each regime, the Convergence Measure would be calculated multiple times, in weak and strong forcing scenarios to ascertain the ensemble size needed in each. It is likely that the strong forcing rain distribution would have more defined (denser) tails, leading to it having a smaller sampling uncertainty for any given ensemble size and as such require a smaller ensemble size than for the weak forcing to reach the same level of sampling uncertainty. This framework depends on the statistic of interest, model variable distribution and desired level of sampling uncertainty. From this study it has additionally been found that since weak and strong forcing regimes have different distribution shapes, their sampling uncertainty convergence with ensemble size will be different. Assuming that other weather regimes will also have distribution shapes specific to that regime, it can be concluded that a different size of ensemble will likely be necessary depending on the weather regime.

The idealised nature of the extended idealised ensemble developed in this paper brings with it limitations. Despite the distributions from the idealised ensemble being compared with ICON-D2 and these being deemed to be realistic enough, they were not exactly equivalent. In addition, the way the model was extended was acceptable for this study however it does not perfectly replicate weak and strong forcing. Particularly because in the weak forcing case, the synoptic forcing term was zero so that when convection did occur, the clouds would not be unrealistically large. However in the real atmosphere, synoptic forcing can still be acting when there is weak forcing. Another limitation is that other sources of uncertainty in ensemble predictions, including model uncertainty and initial condition uncertainty resulting from limited observations or approximations in the DA system, will limit the accuracy of probabilistic forecasts regardless of ensemble size, and was not considered in this study.

This paper provides the starting point for an analysis of how many members are required in operational ensembles based on the (convective) weather regime. The spatial variation of sampling uncertainty is additionally of interest, for

which ICON-D2 simulations are being analysed for the same weak and strong forcing days of Figure 1 (Puh et al., 2023b). Further studies could look at how computational resources are allocated in a forecasting system depending on the convective regimes often incurred. For example in the U.K., 85% of convection is in the strong forcing regime (Flack et al., 2016), compared to 66% over Germany (Zimmer et al., 2011). This would suggest that an ensemble size suited for strong forcing would be of more interest for simulations over the U.K. than for Germany.

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.

Craig, G. C., Puh, M., Keil, C., Tempest, K., Necker, T., Ruiz, J., Weissmann, M. and Miyoshi, T. (2022) Distributions and convergence of forecast variables in a 1000 member convection-permitting ensemble. *Quarterly Journal of the Royal Meteorological Society*.

Done, J., Craig, G., Gray, S., Clark, P. A. and Gray, M. (2006) Mesoscale simulations of organized convection: Importance of convective equilibrium. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **132**, 737–756.

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`.

Flack, D. L., Plant, R. S., Gray, S. L., Lean, H. W., Keil, C. and Craig, G. C. (2016) Characterisation of convective regimes over the british isles. *Quarterly Journal of the Royal Meteorological Society*, **142**, 1541–1553.

Hirt, M., Craig, G. C., Schäfer, S. A., Savre, J. and Heinze, R. (2020) Cold-pool-driven convective initiation: Using causal graph analysis to determine what convection-permitting models are missing. *Quarterly Journal of the Royal Meteorological Society*, **146**, 2205–2227.

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.

Keil, C., Baur, F., Bachmann, K., Rasp, S., Schneider, L. and Barthlott, C. (2019) Relative contribution of soil moisture, boundary-layer and microphysical perturbations on convective predictability in different weather regimes. *Quarterly Journal of the Royal Meteorological Society*, **145**, 3102–3115.

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

Puh, M., Keil, C., Gebhardt, C., Marsigli, C., Hirt, M., Jakub, F. and Craig, G. C. (2023a) Physically based stochastic perturbations improve high-resolution forecast of convection. *Under review*.

Puh, M., Tempest, K. I., Keil, C. and Craig, G. C. (2023b) Flow-dependent representation of forecast uncertainty in a big convection-permitting ensemble. *To be submitted*.

Tempest, K. I., Craig, G. C. and 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**, 677–702. URL: `https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4410`.

Walker, A. (1968) A note on the asymptotic distribution of sample quantiles. *Journal of the Royal Statistical Society: Series B (Methodological)*, **30**, 570–575.

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

Zimmer, M., Craig, G., Keil, C. and Wernli, H. (2011) Classification of precipitation events with a convective response timescale and their forecasting characteristics. *Geophysical Research Letters*, **38**.

6 | ABBREVIATIONS

EnKF ensemble Kalman filter

NWP Numerical Weather Prediction

CLT Central Limit Theorem

DA Data Assimilation

CI Confidence Interval

CLT Central Limit Theorem

CAPE Convective Available Potential Energy

CIN Convective Inhibition

7 | FIGURES AND TABLES

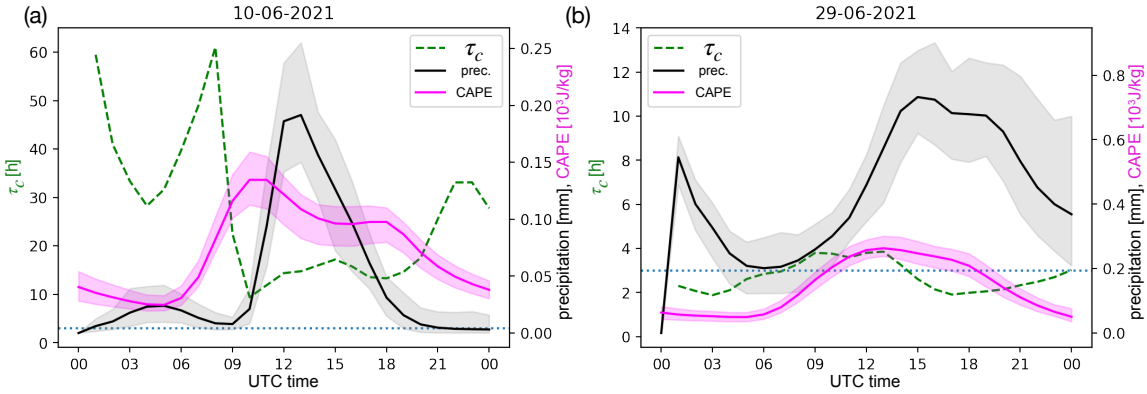

FIGURE 1 Daily evolution of the CAPE, convective timescale (τ_c) and precipitation on the (a) 10th and (b) 29th June 2021 during a period of weak and strong forcing respectively. Shading indicates the 95th CI. Simulated using data from 120-member ICON-D2 ensembles (Puh et al., 2023b).

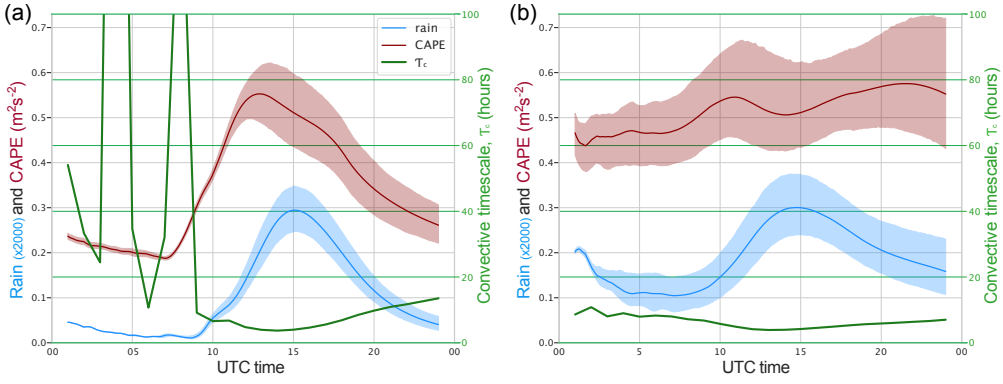

FIGURE 2 Time evolution of total rain in domain, CAPE and convective timescale in the (a) weak and (b) strong forcing free runs from the extended idealised ensemble. Shading indicates the 95% CI spread from the 5,000-member ensemble. A simple moving average is used for the convective timescale.

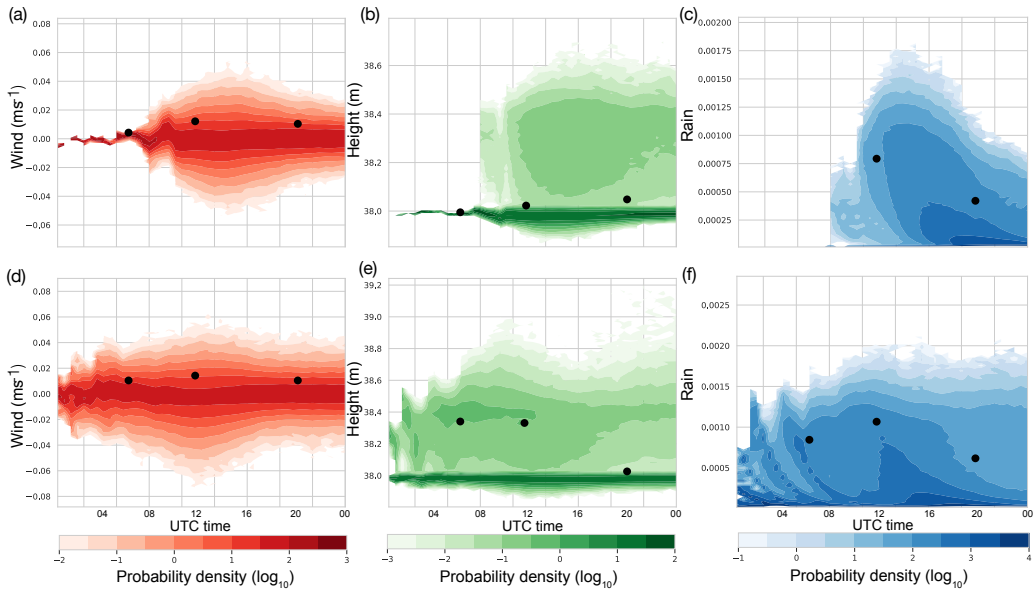

FIGURE 3 Contour histogram plots for the (a,d) wind, (b,e) height and (c,f) rain from the extended idealised ensemble over 24 hours in (a,b,c) weakly and (d,e,f) strongly forced regimes. Black dots indicate location of the 0.95 quantile. Neighbourhood of length 20km shown.

Forcing regime	Time (UTC)	Wind (ms^{-1})	Height (m)	Rain
Weak	06:00	0.000804	0.00169	0
	12:00	0.00808	0.0695	0.000762
	20:00	0.00573	0.0719	0.000138
Strong	06:00	0.00735	0.116	0.000271
	12:00	0.0110	0.121	0.000369
	20:00	0.00722	0.0807	0.000216

TABLE 1 Standard deviation of distributions from Figure 3.

Forcing regime	Time (UTC)	Wind	Height	Rain
Weak	06:00	0.0129	0.00646	0
	12:00	0.0136	0.00262	0.00183
	20:00	0.0134	0.00370	0.00292
Strong	06:00	0.0136	0.00264	0.00526
	12:00	0.0109	0.00374	0.00590
	20:00	0.0135	0.0859	0.00165

TABLE 2 Probability density at the 0.95 quantile of the distributions in Figure 3.

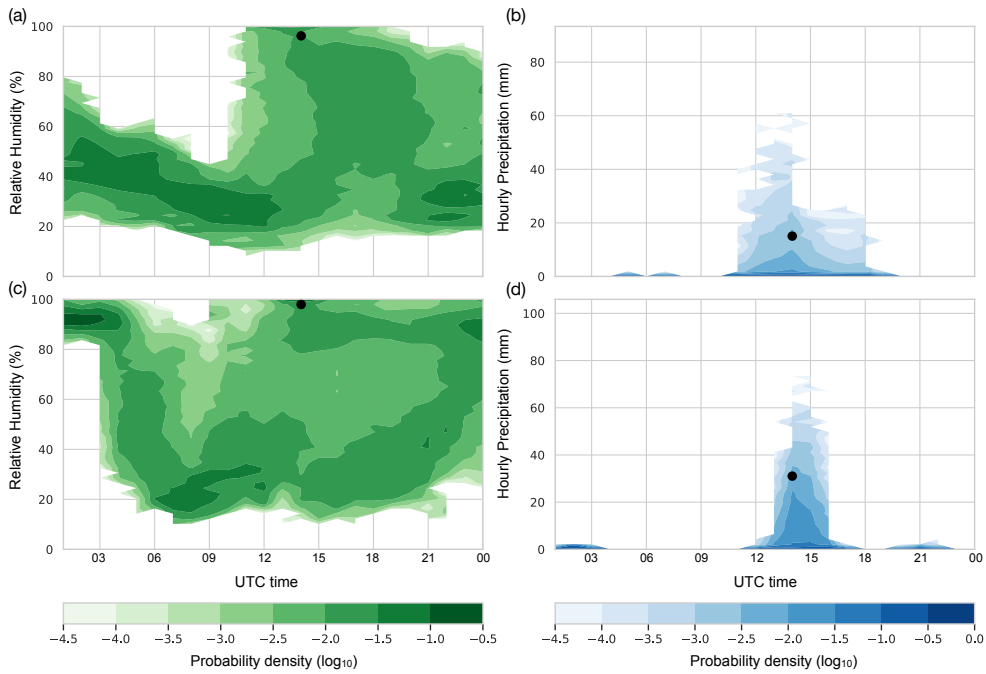

FIGURE 4 Contour histogram plots for the (a,c) relative humidity and the (b,d) hourly precipitation from a 120-member ICON-D2 ensemble over 24 hours on the (a,b) weakly and (c,d) strongly forced regime days of 10 and 29 June 2021 respectively. Data used a circular neighbourhood of radius 10km. Black dots show location of 0.95 quantile at 14:00 UTC.

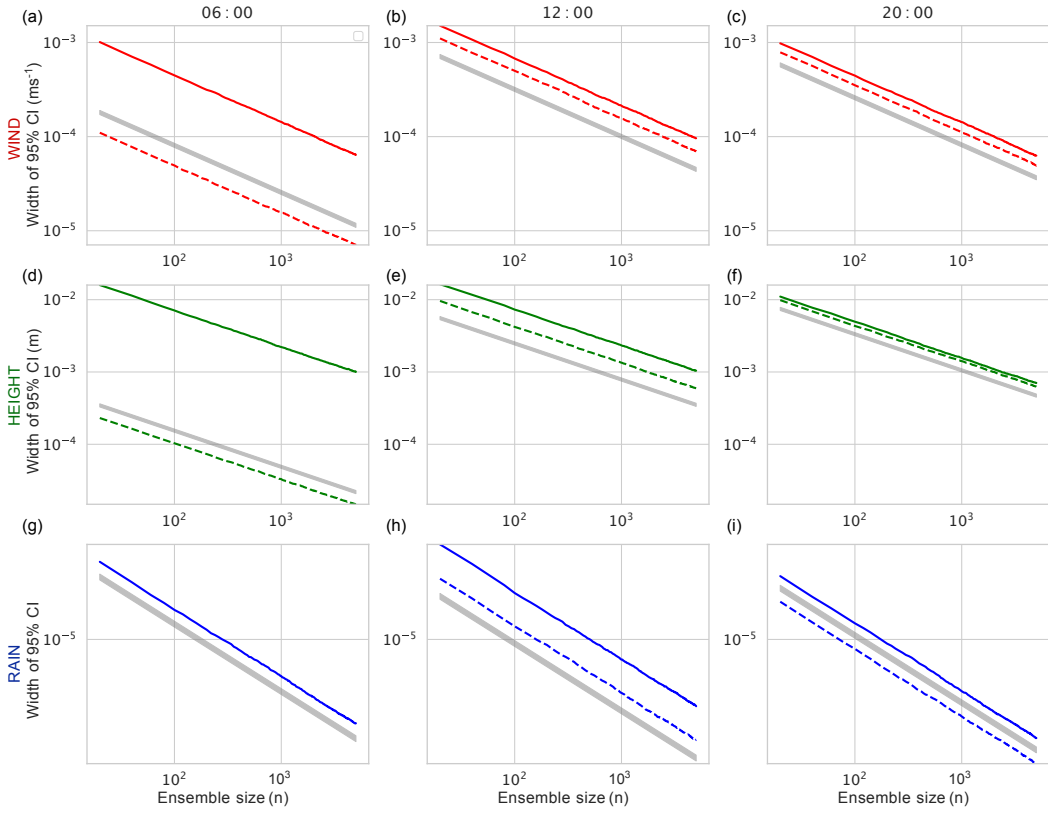

FIGURE 5 Convergence Measure for the mean using data from the extended idealised ensemble with a neighbourhood of length 20km. Shows wind (a,b,c), height (d,e,f) and rain (g,h,i). Each column is a different time point. Weak forcing is indicated by the dashed coloured lines and strong forcing by the solid coloured lines. Grey line in background is converging proportional to $n^{-\frac{1}{2}}$, its width spanning 10% of the magnitude of the weak forcing (strong forcing for (g)) Convergence Measure.

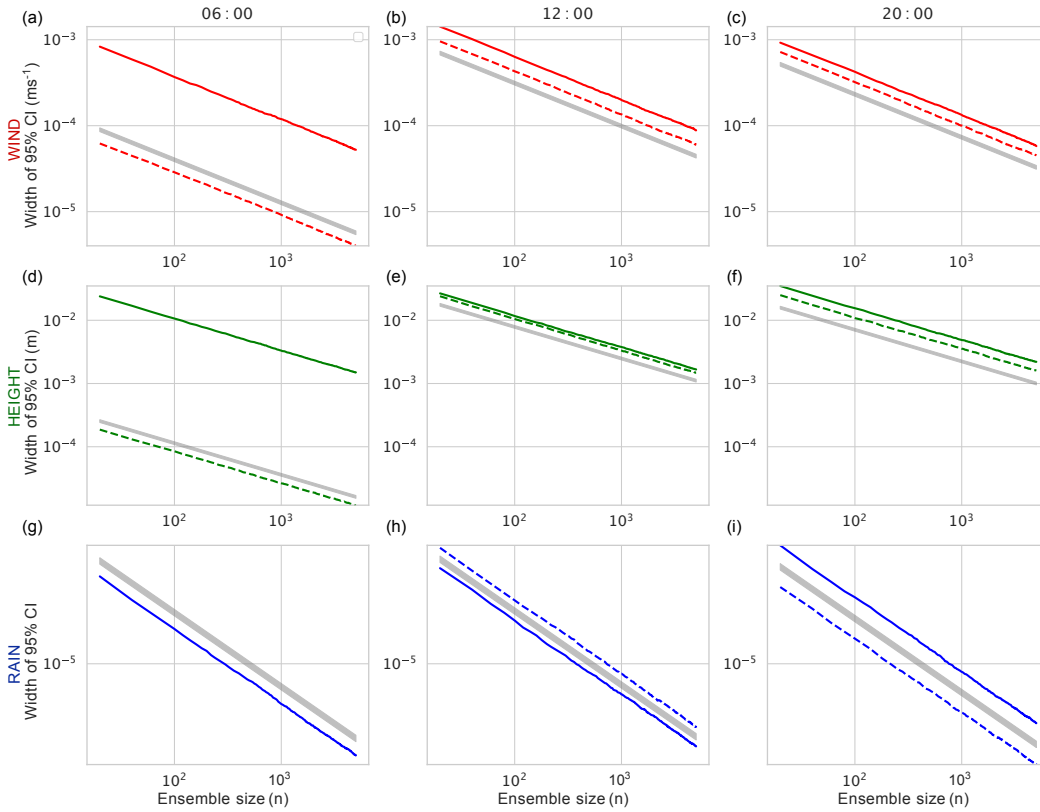

FIGURE 6 As in Figure 5 but for the standard deviation.

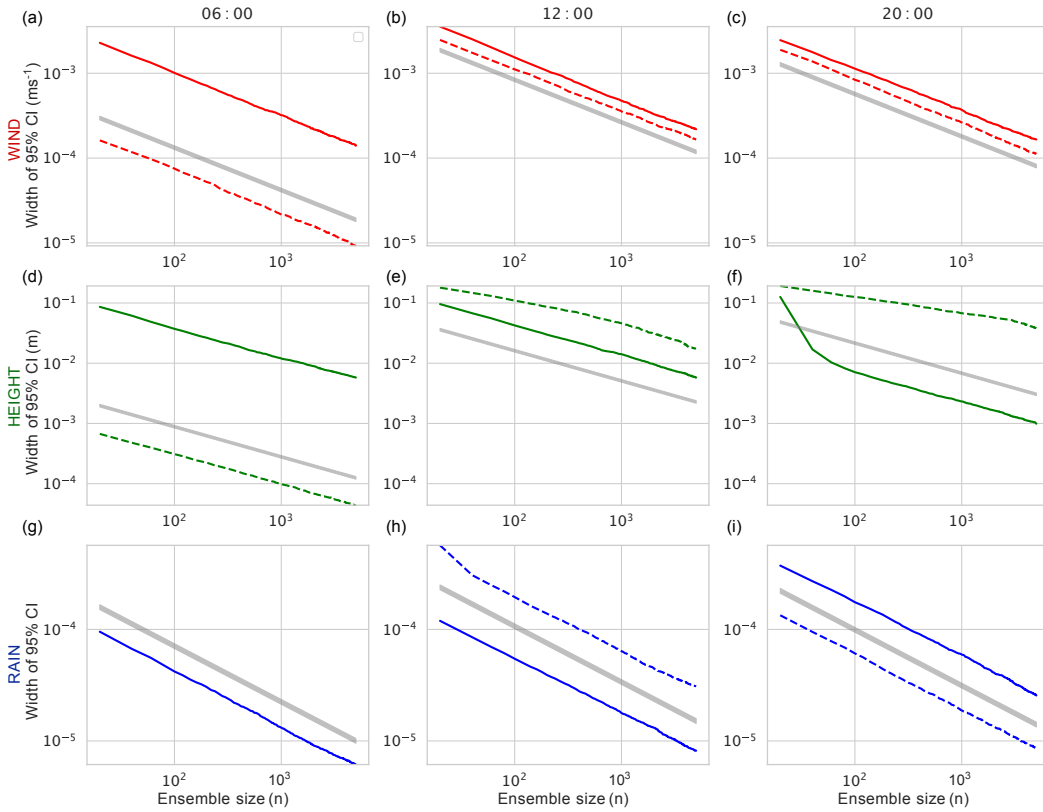

FIGURE 7 As in Figure 5 but for the 0.95 quantile.

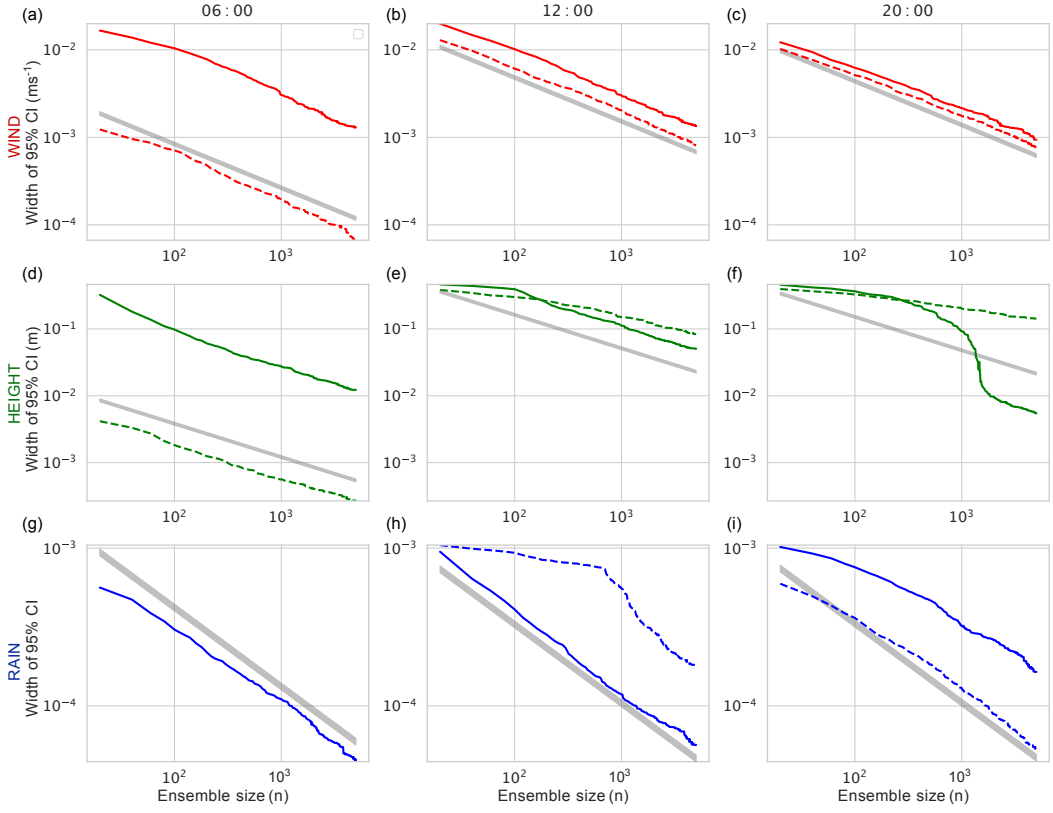

FIGURE 8 Convergence Measure for the 0.95 quantile using data from the extended idealised ensemble from a single grid point. Shows wind (a,b,c), height (d,e,f) and rain (g,h,i). Each column is a different time point. Weak forcing is in the dashed coloured lines and strong forcing is in the solid coloured lines. Grey line in background is converging proportional to $n^{-\frac{1}{2}}$, it's width spanning 10% of the magnitude of the weak forcing (strong forcing for (g)) Convergence Measure.