Sampling uncertainty in ensemble forecasting: When do we have enough ensemble members?

Kirsten Tempest

Munich 2023

Sampling uncertainty in ensemble forecasting: When do we have enough ensemble members?

Kirsten Tempest

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Kirsten Tempest
aus Glasgow, United Kingdom

München, den Juni 1, 2023

Erstgutachter: Prof. Dr. George C. Craig
Zweitgutachter: Prof. Dr. Hella Garny
Tag der mündlichen Prüfung: 10/07/2023

Parts of this thesis are included in:

Tempest, K. I., G. C. Craig, and J. R. Brehmer, 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, doi: https://doi.org/10.1002/qj.4410.
Tempest, K. I., G. C. Craig, M. Puh and C. Keil, 2023: Convergence of forecast distributions in weak and strong forcing convective weather regimes. *To be submitted.*
Tempest, K. I., G. C. Craig, and T. Selz, 2023: Convergence of forecast distributions on the synoptic scale. *To be submitted.*

Abstract

Computing power in weather forecasting is a constraint due to the large number of degrees of freedom in the atmosphere and the limited resources with which to predict it. Nowadays, ensemble forecasting further exacerbates this: In order to get a probabilistic prediction, multiple forecast models are run parallel to each other in an ensemble. Our resources must therefore be used where it is most impactful. One area where it is not clear how many resources are needed is in the size of the ensemble. Sampling uncertainty, which occurs from being unable to sample all the degrees of freedom in the atmosphere leads to imprecise forecasts and can be detrimental in times when it is vital to forecast extreme weather events, especially as climate change makes these more frequent. On the other hand, ensembles may be unnecessarily large for forecasters interested in only basic forecast quantities. By understanding what ensemble size is required in an ensemble forecast, one can better predict the weather as well as ensure computational resources are being optimally spent. An idealised approach is taken, whereby a massive ensemble representative of convective-scale ensemble forecasts is simulated with the aim of looking at how sampling uncertainty decreases with ensemble size. For this, the convergence measure is created which uses bootstrapping (sampling with replacement) of the forecast distribution to obtain a Confidence Interval (CI) within which the sampling uncertainty for a given statistic lies for the sampled ensemble sizes. An asymptotic power law scaling inversely to the square-root of the ensemble size is found in the limit of large ensemble size as a consequence of the Central Limit Theorem (CLT). This is already seen for the mean and variance with operational ensemble sizes. A framework to find the optimal ensemble size given a desired level of sampling uncertainty is then possible if one is in this asymptotic regime by extrapolating the power law to smaller sampling uncertainty levels. If one is not in the asymptotic regime because of too small an ensemble size, a parametric technique which makes use of the distinctive shapes of the forecast variables can be employed. It was furthermore found that the convergence of sampling uncertainty depended on the forecast variable's distribution shape and the statistic of interest. Extending the initial version of the idealised model to include weak and strong forcing convective weather regimes, it is seen that different characteristics of sampling uncertainty exist for different convective weather regimes. This is a consequence of the distribution shapes being different for each regime. The question of ensemble size is not only relevant for convective-scale forecasting, but also for the synoptic scale. By applying the convergence measure to data from the European Centre for Medium-range Weather Forecasts (ECMWF), asymptotic scaling was

confirmed to also occur in synoptic-scale data for statistics including the Extreme Forecast Index (EFI), which is used operationally. A framework for finding the optimal size of ensemble based on the forecast variable, the statistic of interest and the level of sampling uncertainty desired, which is applicable for the convective as well as synoptic scale, is thereby developed in this thesis.

Zusammenfassung

Die verfügbare Rechenleistung ist ein limitierender Faktor in der Wettervorhersage, was vor allem auf die vielen Freiheitsgrade der Atmosphäre zurück zu führen ist. Ensemble-Vorhersagen, die in jüngster Zeit verstärkt durchgeführt werden, verschärfen das Problem: Um eine probabilistische Vorhersage zu erhalten, werden viele, etwas unterschiedliche Simulationen parallel gerechnet, die dann das Ensemble bilden. Die verfügbaren Ressourcen müssen daher möglichst geschickt aufgeteilt werden, wobei allerdings ist die Bestimmung der optimalen Größe des Ensembles schwierig und weitgehend ungeklärt ist. Die Stichprobenunsicherheit, die dadurch entsteht, dass nicht alle Freiheitsgrade der Atmosphäre berücksichtigt werden können, führt zu ungenauen Vorhersagen und dies ist besonders kritisch bei Extremwetter-Ereignissen, welche durch den Klimawandel immer häufiger zu erwarten sind. Auf der anderen Seite sind die Ensembles möglicherweise unnötig groß für die Vorhersage grundlegender meteorologischer Variablen. Ein besseres Verständnis der Anforderungen an die Ensemblegröße einer Vorhersage würde somit bessere Vorhersagen durch eine optimierte Aufteilung der verwendeten Ressourcen ermöglichen. Mit einem idealisierten Verfahren wurde zunächst ein sehr großes Ensemble atmosphärischer Konvektion simuliert und untersucht, wie die Stichprobenunsicherheit mit der Ensemblegröße abnimmt. Dazu wurde eine Konvergenzmetrik entwickelt, welche die Bootstrapping-Technik (mit Zurücklegen) auf die Vorhersageverteilung anwendet, um ein Konfidenzintervall zu erhalten, in welchem die Stichprobenunsicherheit für eine bestimmte statistische Größe und für gegebene Ensemblegröße liegt. Asymptotisch zeigte sich ein Potenzgesetz invers zur Wurzel der Ensemblegröße im Grenzfall großer Ensembles als Konsequenz des zentralen Grenzwertsatzes. Für den Mittelwert und die Varianz ist dieser Grenzfall für operationelle Ensemblegrößen bereits erfüllt. Durch Extrapolation des Potenzgesetzes ist es dann in dem asymptotischen Regime möglich, die optimale Ensemblegröße für eine beliebige gewünschte Stichprobenunsicherheit zu bestimmen. Außerhalb des asymptotischen Regimes, d.h. bei zu kleiner Ensemblegröße, kann eine parametrische Technik angewendet werden, welche die charakteristischen Verteilungen der Vorhersagevariablen ausnutzt. Weiter konnte gezeigt werden, dass die Konvergenz der Stichprobenunsicherheit von der Art der Verteilung der vorhergesagten Variable und der gewünschten Statistik abhängt. Mit einer Erweiterung des idealisierten Modells auf jeweils schwach oder stark angetriebene Konvektion konnte außerdem gezeigt werden, dass sich die Charakteristiken der Stichprobenunsicherheit abhängig vom konvektiven Wetterregime unterscheiden. Dies ist eine Konsequenz der unterschiedlichen Verteilungsformen in den beiden Wetterregimen. Die

Frage nach der Ensemblegröße ist jedoch nicht nur für Konvektion von Interesse, sondern auch für größere, synoptisch-skalige Prozesse. Dazu wurde die Konvergenzmetrik auf Vorhersagedaten des European Centre for Medium-range Weather Forecasts (ECMWF) angewendet. Die asymptotische Skalierung konnte auch für die synoptisch-skaligen Daten für verschiedene statistische Größen bestätigt werden, einschließlich des operationell verwendeten Extreme Forecast Index (EFI). Somit stellt diese Arbeit ein Verfahren vor, um die optimale Ensemblegröße in Abhängigkeit von der vorhergesagten Variable, der gewünschten Statistik und der gewünschten Stichprobenunsicherheit zu finden und das sowohl auf der konvektiven als auch auf der synoptischen Skala.

Contents

Chapter 1

Introduction

Ensemble forecasting, whereby forecast models with slightly different initial conditions are run in parallel, has become progressively more commonplace in operational weather centres worldwide since the 1990s. Evolving from single, deterministic, weather models, ensemble forecasting allows for a probability to be attached to a meteorological prediction, for example to say that there is a 40% chance of rain. The improvement in weather forecasting has been, and continues to be, hugely important for a strong economy and in protecting human life and property (Craig et al., 2021)[1], especially since humans must become more resilient with the effects of climate change creating more frequent natural disasters such as wind storms and flooding. To emphasise the positive impact improving forecasts can have, the World Bank has estimated that there could be increases of up to 30 billion USD per year in global productivity and 2 billion USD per year in reduced asset losses, from investing about 500 million USD in improving weather, climate, and water observation and forecasting systems (Anderson et al., 2017). This provides a great motivation for further improving weather forecasts, which is the overarching goal of this thesis.

1.1 History and Outlook of Weather Forecasting

In the following I will outline the progression from deterministic forecasting to probabilistic predictions using ensembles of deterministic forecast models. Finally, I will conclude with the vital questions being asked today in the field of ensemble forecasting, focusing on the question of ensemble size which I will explore in more depth.

1.1.1 Deterministic forecasting

Abe (1901) and Bjerknes (1904) saw the potential of applying the laws of physics to the problem of predicting the weather in the early 1900s. By using differential equations to

[1] I am a co-author, whereby I contributed the sub-section “Early-career scientists”

calculate how variables of the atmosphere, namely wind, pressure, density and temperature, change in time, future states of the atmosphere could be estimated. These equations were composed of the Navier-Stokes and mass continuity equation as well as the law of thermodynamics (Bauer et al., 2015). The differential equations were then solved numerically using spatial and temporal discretisation. The first successful computation of these equations was carried out by Charney (Charney, 1948) in the middle of the 20th century. A few years following this, the first real-time deterministic forecast was created in Stockholm (Bolin, 1955).

It happened that deterministic forecasting, whereby there is only one possible state of the atmosphere predicted, was often misleading. For example in October 1987 the deterministic forecast currently in operational use at that time did not foresee a storm over the UK. Had an ensemble been run however, there may have been an extreme weather warning broadcasted as many members from a later simulated 51 member ensemble showed a deep depression (low pressure) and strong winds, not shown in the deterministic forecast at the time (Slingo and Palmer, 2011). From this and other similar incidents, it was made clear to forecasters and researchers that deterministic forecasts weren't the ultimate answer to weather forecasting but rather an important stepping stone along the way. Slowly, the world of operational forecasting was beginning to delve into the possible advantages of a probabilistic forecast.

1.1.2 Ensemble forecasting

In the second half of the 20th century scientists were investigating the limits of predictability of the atmosphere, that is, how far in advance they could provide a prediction which would be better than a random one. Limits on the predictability of the atmosphere exist because it is very chaotic, with $10^6 - 10^8$ degrees of freedom (Leutbecher and Palmer, 2008). In order to first understand this chaos, the divergence of initial conditions in non-linear systems was investigated. One of the first experiments to address this calculated the Root Mean Square Error (RMSE) (measures the variability) of wind error predictions. In this experiment the wind error was seen to double in two days (Thomson, 1957). Lorenz (1963) quantified this result further by using a finite system of deterministic non-linear differential equations. He discovered how predictability is flow dependent and how two slightly different initial conditions can lead to remarkably different states of the atmosphere at a later point in time, the so called “Butterfly effect” (although he references a seagull's wings rather than those of a butterfly's in his work). Lorenz argued that deterministic forecasts could not be trusted due to the Butterfly effect and the ability of tiny errors to have a large impact on the forecast. As such he was one to foresee the future ensemble system (Lorenz, 1965).

Multiple methods have been proposed to obtain a probabilistic future state of the atmosphere. This has included the Liouville or Fokker-Planck equations (Palmer, 2017) and Epstein's stochastic dynamic prediction method (Epstein, 1969). Both were limited however by the large number of degrees of freedom of the atmosphere (Leutbecher and Palmer, 2008). The Monte-Carlo method was ultimately suggested (Leith, 1974), which is used to this present day, being operational for over 25 years (Palmer, 2019).

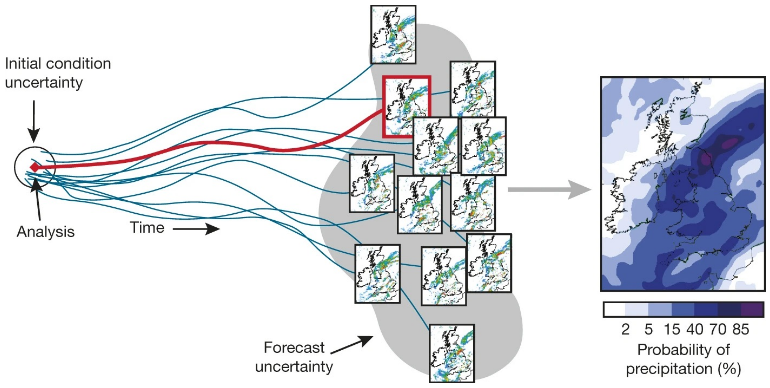
Initial condition uncertainty
Analysis
Time
Forecast uncertainty
2 5 15 40 70 85
Probability of precipitation (%)

Figure 1.1: Example of ensemble forecasting. Initial condition uncertainty is propagated in time with an ensemble of predictions (shown by blue lines) to sample the forecast uncertainty at a later time point. The red line shows the analysis being propagated in time. This set-up enables a probability to be attached to the meteorological prediction, in this case precipitation. Figure originally from Bauer et al. (2015).

The Monte-Carlo method commonly used in ensemble forecasting begins with a distribution of states of the atmosphere, which is provided by Data Assimilation (DA). Data assimilation simply combines the prediction from a numerical forecast, known as the background, with observations of the atmosphere to calculate a best-guess of the current state of the atmosphere, known as the analysis. Alongside this an associated error is calculated, which is known as the initial condition uncertainty and can be used to create a distribution of initial states for the Monte-Carlo forecast. Each of these states is known as an ensemble member and together they make up the ensemble. The ensemble distribution therefore is made up of ensemble members from a specific forecast variable. The initial condition uncertainty is seen in Figure 1.1, whereby the ensemble members, shown by blue lines, are propagated in time by forecast models which have only slight differences between them to allow for forecast spread. The analysis ensemble member is shown in red and the grey shading is the forecast uncertainty at a later time which has been calculated from the resulting projections from each of the ensemble members. This allows for a probabilistic prediction of the forecast variable, which in this case is precipitation. The forecast uncertainty comes from multiple sources of unpredictability including small scale noise, upward propagation of energy, errors from numerical and physical approximations and inaccurate and lack of observations to sample the initial state of the atmosphere.

From constant improvements in a multitude of areas including from advancing the

DA (e.g. Ruckstuhl and Janjić (2018)), observation systems (e.g. Wang et al. (2015)) and the formulation of the numerical models (e.g. Hirt et al. (2019)), we've been able to extend the forecast by roughly one day each decade for the past 40 years (Bauer et al., 2015). Further improvement is possible, as shown by Selz et al. (2022) who calculated that 4-5 days of predictability can be gained by continued improvements to the forecasting system. According to Bauer et al. (2015), there are three main areas where improvement is needed in ensemble forecasting. These are the representation of physical processes in the ensemble models, the initialisation of the model and the ensemble forecasting itself. This thesis concentrates on improvement in the latter area, with a specific focus on the question of how to decide on how many ensemble members are required for the ensemble forecast.

1.1.3 Ensemble Size

Within the ensemble forecasting system, the number of ensemble members used in operational centres has not changed significantly in recent years. Operational ensembles typically have sizes between 20 and 50 members (Buizza et al., 2000; Reinert et al., 2020; Metoffice, 2023). These sizes have often been chosen based on the computational power available at that time. As it is important that forecasts are improved, and ensemble forecasting is one of the key areas to improve, it is consequential to ask how many members are actually needed in an ensemble.

It is not precisely known what ensemble size is optimal to have in weather forecasting. Leith (1974) suggested from looking at the Mean Square Error (MSE) that eight members would be enough to estimate an ensemble mean and not much improvement would be gained from adding further members. This was, however, calculating the mean of the ensemble members and not statistics which forecast extreme weather events, which are often difficult to accurately forecast with current operational ensemble sizes. Extreme weather events are events such as major floodings from a lot of precipitation or high speed winds, where the probability, and as such the predictability, of the event occurring is low. Extreme weather events is one of the top two risks facing the world today and so it is a priority to forecast them reliably (WEF, 2020). In order to predict extreme events, the probability of extreme events occurring needs to be known and the quantiles of tails of distributions are a statistic which is often used for their prediction. Lovejoy and Schertzer (2018) show however that tails of uni-modal distributions are not well resolved with current ensemble sizes of about 50. This sensitivity for quantiles where the probability density is low, which will be referred to as extreme quantiles, is perhaps unsurprising considering that the frequency distribution of a forecast quantity (hereafter referred to simply as distribution) from an ensemble of up to 50 members is unlikely to be accurate for rare events that are infrequently sampled. Whereas fewer members would be needed to resolve a less extreme (e.g. median) quantile. Clearly forecasters are interested in different statistical properties of the forecast, and as such there will likely not be one magic number for how large an ensemble should be. Rather, it will likely depend on what is interesting to the forecaster, as well as the forecasting system itself.

It has been seen from many studies that increasing ensemble size does have a positive effect on the quality of the forecast. Buizza et al. (1998) and Raynaud and Bouttier (2017) compare the benefits of increased ensemble size against those of higher resolution, where the grid points in the model are closer together, for global and regional forecasting systems. Both studies found that either ensemble size or resolution increases could be more beneficial, depending on factors such as forecast lead time (duration of forecast run) and the quantity being predicted. Richardson (2001) took into account the usefulness of the forecast for users. They showed that users with low predictability forecasts and low cost/loss ratios (where the cost of taking preventive action is much less than the cost of what could be lost if preventive action is not taken) would particularly benefit from increasing their ensemble size. Machete and Smith (2016) took another approach by calculating an ensemble's relative information content (how much information the ensemble contains) to measure the effect of increasing the ensemble size. They showed that there is still information to be gained from increasing ensemble size even when ensembles are on the order of 100 in size, however at a large enough ensemble size model error will eventually dominate and sampling uncertainty won't be of concern. There are studies which showed no significant effect of increasing ensemble size e.g. Bannister et al. (2017); Jirak et al. (2016). In both of these studies however the maximum ensemble size was only 93. In general, from looking at the impact of ensemble size in specific instances and scenarios it appears that ensemble size does have a measurable effect on the quality of the forecast. This shows that understanding the ensemble size required is an important issue. It is not clear though from these studies what a large enough ensemble size would be. From the dependencies on specific cases and models, it is likely that a general framework is needed to understand this in detail.

Steps to understand ensemble size required have been taken from a theoretical standpoint. Leutbecher (2019) delved into the problem of how large an ensemble should be by providing 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 (if the forecast model is able to replicate the observed state) and that the members are exchangeable (the members could be mixed up and it would not make a difference). Under these assumptions for an ensemble of size n, the score of the Continuous Ranked Probability Score (CRPS), which measures the performance of an ensemble, is equal to the score for an infinite ensemble multiplied by a factor $(1+\frac{1}{n})$. This shows that improvements in CRPS will be small once the ensemble size has reached a few 10s of members, and useful estimates could often be obtained with even smaller ensembles. Similar results were found for other scores, with the notable exception of the Quantile Score (QS) which evaluates how well quantiles of value $0 < p < 1$ can be measured in a forecast (Leutbecher, 2019). For the more extreme quantiles on a uni-modal distribution close to 0 or 1, convergence required much larger ensemble sizes. These results encourage the formulation of a broad framework for understanding and estimating ensemble size, however there needs to be further refinement for it to be applicable to every ensemble forecast and statistical quantity.

In research environments, larger ensemble sizes have been considered to resolve the forecast distribution more accurately and to investigate the question of ensemble size. For example, Lin et al. (2020) evaluated a measure of hurricane strength called non-dimensional damage, that depends non-linearly 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, which quantifies how different forecast variables are dependent on each other and is used in DA. This was found to be of benefit since multivariate combinations of values may be infrequently sampled even when the individual values are not rare. This effect is magnified from simple Gaussian marginal distributions being able to create more complicated multivariate distributions (Poterjoy, 2022). 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), to measure the degree of non-Gaussianity, how much the forecast distributions diverged from being Gaussian distributed, at different ensemble sizes. It was found that in general, approximately $1,000$ members were required to represent characteristics of non-Gaussian distributions such as skewness and kurtosis, the third and fourth moments of the distribution. Using the same model as Miyoshi et al. (2014), Necker et al. (2020a) quantified how sampling uncertainty decreased in spatial covariances of smaller subsets of their $1,000$-member simulation over a domain of central Europe. Furthermore in Necker et al. (2020b), a look up table approach was developed to correct for sampling uncertainty in the spatial covariances which was dependent on ensemble size. These studies have often come to the conclusion that a $1,000$-member ensemble should be used. It is not clear however how this number should differ depending on the ensemble forecast, or the forecasting case.

Due to the large degrees of freedom in the atmosphere, ensembles of operational size contain a measurable sampling uncertainty. This occurs because an ensemble can't create a perfect replica of the real distribution when there are less ensemble members than degrees of freedom. This sampling uncertainty then leads to inaccurate forecast predictions. To estimate the extent of this inaccuracy, a form of bootstrapping can be used (Davison and Hinkley, 1997). Bootstrapping is simply sampling from a distribution with replacement, to create a statistically identical new distribution. Furthermore, non-parametric bootstrapping is when the underlying distribution is not known. Non-parametric bootstrapping can then be used to infer statistical properties about the underlying distribution without making assumptions about it. It samples with replacement from an empirical Cumulative Distribution Function (CDF) to create bootstrapped distributions which can then be used to create Confidence Interval (CI)s which provides a probability with which a statistic is within the limits of the chosen interval (e.g. 95%) (Jolliffe, 2007). This can be useful in estimating the actual sampling uncertainty. Dibike et al. (2008) used this non-parametric bootstrapping method to quantitatively evaluate the uncertainty of statistically down-scaled climate data in Northern Canada, whereby the low resolution climate data was processed to obtain higher resolutions. To calculate the uncertainty, they constructed CIs to calculate the variability of the mean and spread of the difference between the down-scaled and observed, meteorological variables. In addition, Feng et al. (2011) used block bootstrapping to predict the uncertainty related to seasonal means. In their case, blocks of data were sampled rather than individual data points so as to keep their serial time correlation. From these studies it is clear that bootstrapping can be a powerful tool in calculating the magnitude of the sampling uncertainty created from having a finite ensemble size.

The previous studies detailed the strong impact ensemble size can have on forecast skill and sampling uncertainty as well as first steps to estimate what ensemble size would be optimal. Despite this, it is not clear how many members are required to resolve the full distribution including the tails and forecast them accurately enough. It is likely a framework is needed, that can be applied to a specific ensemble and forecast case, to reach an ensemble size relevant to the forecaster. This leads to the central question of this thesis which is **how to know what ensemble size is required to achieve sufficient accuracy in your statistic of interest?**

1.2 Asymptotic Convergence of Sampling Uncertainty

In order to address the central question of what ensemble size to aspire to, preliminary studies have looked at the nature of how sampling uncertainty of probabilistic weather forecasts decreases with increasing ensemble size. It is thought that if this converges with ensemble size according to a theory, then it can be estimated how further ensemble size increases would influence the accuracy of the forecast. This has been investigated from an experimental and theoretical point of view, each of which will be explored in this section.

1.2.1 Convergence in forecast data

Milinski et al. (2020) used bootstrapping without replacement to measure how sampling uncertainty decreased with ensemble size. With a 200-member climate model, a power law like convergence was observed for statistics of the global mean and RMSE over a region. Their method of bootstrapping without replacement limited them however in determining how further increases in the ensemble size would reduce the sampling uncertainty. This is because the sampling uncertainty at the maximum ensemble size would be zero, which is very unrealistic. A "recipe" was outlined, explaining how one could determine the ensemble size required based on a maximum level of uncertainty acceptable for a forecaster. If one could estimate how sampling uncertainty would continue to decrease at larger ensemble sizes, one could give a best guess on how many ensemble members they would need to reach their required level of sampling uncertainty.

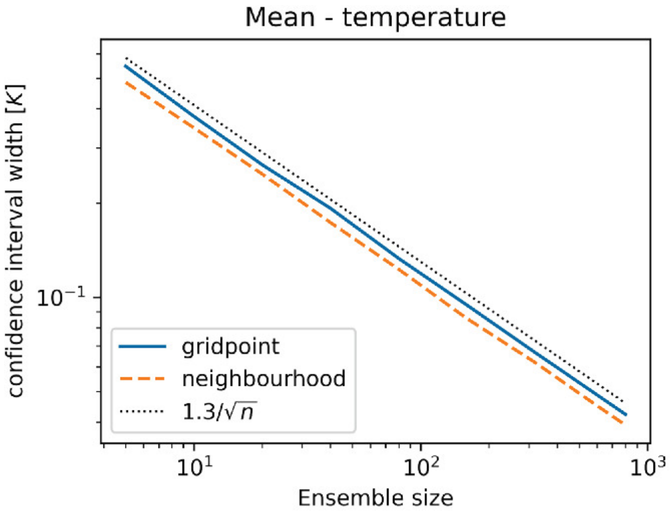

Figure 1.2: Convergence of sampling uncertainty with ensemble size for the mean statistic and the temperature distribution. Log_{10} scaled x and y axis. The orange dashed and blue solid line correspond to two different methods of calculating how sampling uncertainty decreases with ensemble size. The dotted black line scales proportional to $n^{-\frac{1}{2}}$, where n is ensemble size. Figure from Craig et al. (2022).

To avoid the problem of an unrealistic sampling uncertainty measurement at the largest ensemble size as in Milinski et al. (2020), Craig et al. (2022) [2] used bootstrapping with replacement on their 1,000-member ensemble to investigate how sampling uncertainty would decrease with ensemble size. We discovered that for all distribution shapes and most forecast variables, the width of CIs, which is used as the measure of sampling uncertainty of the ensemble estimates, decreased proportional to $n^{-\frac{1}{2}}$ with increasing ensemble size n. As this scaling extended infinitely, not going to zero, it is called “asymptotic” (Urdan, 2022). The forecast variables included the mean, standard deviation and the 95th percentile of temperature and humidity, as well as the probability of precipitation exceeding certain thresholds. An example is given for the mean of the temperature distribution in Figure 1.2. It is seen that using a neighbourhood (orange line), whereby the ensemble size is artificially increased, the sampling uncertainty is smaller than that for the distribution from a single grid point (blue line). We found the $n^{-\frac{1}{2}}$ scaling for sufficiently large ensemble sizes for all statistical quantities, except some 95th percentiles and the probability of precipitation

[2] I was a co-author and aided in the analysis of data and interpretation of results and contributed to writing.

exceeding large thresholds. We hypothesised that the scaling would eventually be observed for these statistical quantities where the scaling had not been observed, if ensembles sizes larger than $1,000$-members would be employed. Additionally, we saw that the distribution shape of the forecast uncertainty can influence convergence of the sampling uncertainty of a statistical quantity.

The two studies (Milinski et al., 2020; Craig et al., 2022) described in this section show there is potential in looking at how sampling uncertainty decreases with ensemble size to answer the central question of how large an ensemble should be. Since if the sampling uncertainty always converges with the scaling of $n^{-\frac{1}{2}}$, one would be able to estimate how small the sampling uncertainty would become with even larger ensemble sizes. Craig et al. (2022) also pointed to the potential importance of distribution shape of the forecast uncertainty on the convergence behaviour.

Distribution shapes and neighbourhood method

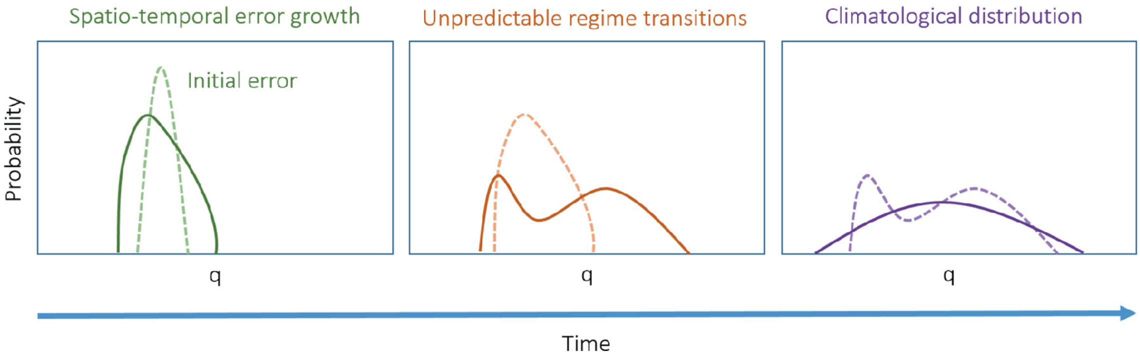
Spatio-temporal error growth
Unpredictable regime transitions
Climatological distribution
Initial error
Probability
q
q
q
Time

Figure 1.3: Conceptual model of the evolution of forecast uncertainty for a fictitious variable “q” in a forecast. Dotted line is distribution shape from previous time step. Figure from Craig et al. (2022).

As previously mentioned, we hypothesised that the convergence of sampling uncertainty depends on the distribution shape (Craig et al., 2022). It was generally seen that more complicated shapes which were less Gaussian would need a larger ensemble for their convergence to scale proportional to $n^{-\frac{1}{2}}$, and different shapes would lead to different magnitudes of sampling uncertainty. An idea of how forecast distribution shapes vary as a function of lead time is given by our conceptual model, shown in Figure 1.3. It shows the distribution broadening from the constrained initial distribution prepared by the DA, as uncertainty increases and deviations from Gaussianity become larger as non-linear processes (for example convection) become important. These deviations may be of the form of the distribution developing long tails (extreme events) and multiple peaks (weather regimes). At long lead times it will then converge to a smoother climatological distribution. Looking at forecasts of winds, temperatures, humidity and precipitation from a $1,000$-member cloud-resolving ensemble, we found evidence for this progression of distribution shape. As distribution shapes don't remain constant but change and evolve during a forecast run, this means that their effect on the convergence of sampling uncertainty will change throughout a forecast.

A variety of shapes of distribution may be seen in a forecast run. For example, Thomas et al. (2021) show through a large eddy (small scales) simulation and a Gaussian mixing model that the model variable of water vapour saturation density is negatively skewed. Censored, shifted gamma distributions were found to fit precipitation accumulations (Scheuerer and Hamill, 2015) and Jacques and Zawadzki (2015) saw in their convection-resolving model a Laplace distribution fit the horizontal wind, as well as some bimodality occurring. We identified three categories of distributions: quasi-Gaussian, highly skewed and multi-modal (Craig et al., 2022), which have also been observed in other studies, e.g. Kawabata and Ueno (2020). Clearly there are various types of distributions being formed in an ensemble, not only Gaussian distributions which are often assumed by DA algorithms. It will be of interest to look at their impact on the convergence of sampling uncertainty with ensemble size.

Oftentimes the shape of a distribution will not be clear due to too small an ensemble. To counter this, the neighbourhood method can be employed. This method was shown to be effective in our study with a $1,000$-member full convective-scale Numerical Weather Prediction (NWP) ensemble (Craig et al., 2022) where a larger ensemble was needed to see asymptotic convergence in some forecast variables. The convective scale has an order of $\mathcal{O}(10\text{km})$ and the "full" NWP model simply means that the numerical model has a complexity which captures all relevant processes in the atmosphere and is of an operational standard. The neighbourhood method can be used to create smoother distribution shapes from an otherwise small ensemble with a large sampling uncertainty. 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.

Previous studies have found a large array of distribution shapes within their ensembles. This is of potential importance because it has been hypothesised that the ensemble distribution shape can affect the convergence of sampling uncertainty with ensemble size. As such, it will be of interest to look at the shapes of forecast distributions and understand their link to the sampling uncertainty of statistical quantities.

1.2.2 Convergence in theory

Asymptotic convergence of sampling uncertainty, which has been observed in meteorological data, whereby convergence proportional to $n^{-\frac{1}{2}}$ occurs indefinitely in the limit of large n, is a well-established topic in statistics. This exists due to the Central Limit Theorem (CLT). The CLT states that 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 upon the initial distribution's shape (Dekking et al., 2005). It further assumes that the underlying distribution of random variables has a finite variance.

The CLT can be used to calculate a standard error. A standard error is "a measure of variability between samples if an infinite number of samples could be drawn from a population" (Harding et al., 2014). This means that the standard error is not the same as the standard deviation which measures the variability of one particular sample. Rather, it is the standard deviation of the sampling distribution of the statistic of interest and can be referred to as the sampling uncertainty. For a few statistics the standard errors are well-known and can be simply quantified (Harding et al., 2014). Perhaps the most well known standard error is that for the mean, estimated as:

$$\sigma_m = \frac{s_X}{\sqrt{n}}, \tag{1}$$

where s_X is an estimate of the population's standard deviation. It can be seen that the standard error 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 $n^{-\frac{1}{2}}$:

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

although this requires normality of the underlying distribution. Note that these estimates do not address how many members are needed for asymptotic convergence to begin.

Convergence of sampling uncertainty according to Equation (1) was demonstrated in Leutbecher (2019) with an idealised set-up. Multiple simulated Gaussian distributions with up to 16,000-members were used to measure how the uncertainty of the ensemble mean converged with ensemble size. A close match between the measured value of the sampling uncertainty for the mean as well as the variance with the theoretical value for various ensemble sizes was found.

As well as for the mean and standard deviation, a well known equation exists which describes how the sampling uncertainty of any quantile decreases with ensemble size. For a given quantile level, the standard error of the ensemble sample estimate of that quantile is given by:

$$\sigma_p = \frac{1}{\sqrt{n}}\sqrt{\frac{p(1-p)}{f^2(q_p)}}, \tag{3}$$

where 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 (Gneiting, 2014; Stuart and Ord, 2000). The first term on the right-hand side of Equation (3) 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, i.e. predictions of rare events are less confident. For sufficiently large n, Equation (3) 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 1.4, 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 (3). 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 (in this case further away from the median), the sampling uncertainty increases for any given number of ensemble members, varying inversely with the underlying Gaussian distribution shown in Figure 1.4(a).

It has been seen that the power law behaviour of sampling uncertainty convergence with ensemble size observed in computational meteorology studies occurs due to the CLT. This provides a strong basis for creating a framework with which one could estimate the required ensemble size based on the level of sampling uncertainty acceptable, in cases where asymptotic scaling can be observed in forecast data. For example, if a forecaster is wanting to approximate the number of members required to reach a certain accuracy in the spread of their measurement of temperature over Munich they could use asymptotic scaling. This would work by quantifying how their data which they have with their current sized ensemble scales with $n^{-\frac{1}{2}}$, and then extrapolating this until it reaches the level of sampling uncertainty they would wish to have. The corresponding ensemble size required to get that level of sampling uncertainty would then be the desired ensemble size. Previous studies have made a solid start in investigating the possibilities of this technique, however these have been preliminary and further exploration is required. This motivates us to

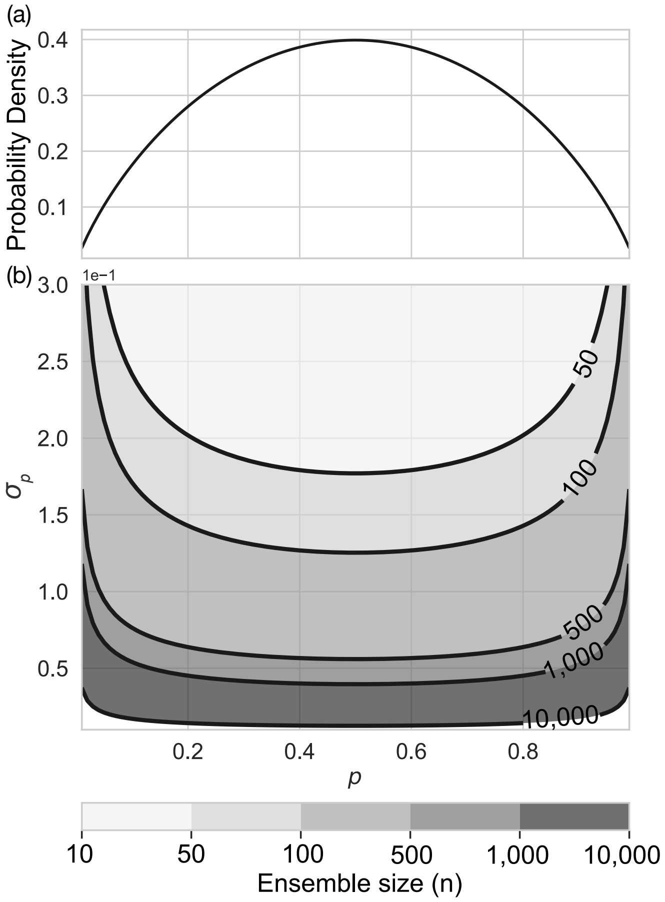

Figure 1.4: (a) Probability density of a Gaussian as a function of quantile p and (b) showing the corresponding ensemble size (contours) required to reach a given level of sampling uncertainty (y-axis) for different quantile levels (x-axis).

thoroughly answer **whether asymptotic theory is relevant to real forecast data, by quantifying how sampling uncertainty decreases with ensemble size**.

1.3 Convergence of Sampling Uncertainty in Convective Weather Regimes

Forecast uncertainty is impacted by the meteorological situation (Keil and Craig, 2011). Weather regimes group specific conditions of the meteorological environment together and convective weather regimes are those regimes involving convection. For example, convection in the midlatitudes can often be categorised by the two regimes of weak and strong forcing (Flack et al., 2016). As each weather regime will produce different distribution shapes of forecast uncertainty for various variables e.g. precipitation, due to the different atmospheric processes, it is of interest to look at how the convergence of sampling uncertainty with ensemble size differs between different regimes. It could be that the ensemble size required is dependent on the weather regime. Due to myself being situated in the midlatitudes, weak and strong forcing convective weather regimes will be investigated in this thesis. As a first step to understand the two regimes, there will be an introduction to convection. Thereafter the two convective weather regimes will be described.

The following introduction to convection and weather regimes is adapted from the textbooks Emanuel (1955) and Lin (2007).

1.3.1 Convection

In atmospheric sciences, convection generally refers to the small-scale thermally-driven circulations which result from gravity acting upon an unstable vertical distribution of mass in the atmosphere. This convection combined with moisture (known as moist convection) organises itself on many spatial scales, producing microscale turbulence to fronts and hurricanes.

The parcel method is used to assess the stability of the atmosphere to convection. By calculating the buoyancy of a parcel of air that is displaced vertically a finite distance, it can be determined whether it will rise, stay where it is, or sink from its current position. If the parcel has water vapour available, it will allow for condensation as the parcel rises. As the parcel ascends and the saturation vapour pressure decreases, this lets more water vapour in the parcel condensate. With this condensation, latent heat is released, warming the parcel and keeping it more buoyant than it's environment when it is displaced upwards. This convective instability with the involvement of water vapour is known as conditional instability and leads to moist convection and precipitating clouds. The point at which the parcel begins to be more buoyant than its environment is known as the Level of Free Convection (LFC). If there is a region below the LFC where the parcel is less buoyant than its environment, it is known as Convective Inhibition (CIN). If the CIN is overcome and passes the LFC, the parcel will keep rising until the surrounding environment is as dense as the parcel, meaning that it is no longer positively buoyant. This occurs at the Level of Neutral Buoyancy (LNB).

There are three main stages to a single convective storm, often thought of as the lifecycle of a cumulonimbus cloud, which occurs as a result of conditional instability. First there is the developing stage, in which there is a warm, strong updraft and air is pulled in at the cloud boundaries (entrainment). This can be catalysed from thermals (surface fluxes of latent and sensible heat) or mountains where air packets are forced to rise. Raindrops and ice particles will begin to form at this stage in the upper portion of the cloud as condensation arises but there will not be substantial rainfall at this point. The second stage is the mature stage. Here the cloud is continuing to grow. Precipitation is now falling and reaching the ground which then evaporates, cooling the base. In the third stage, dissipation begins. Here detrainment (opposite of entrainment) begins at the centre of the cloud, whereby it no longer brings in air from its surroundings. Precipitation meanwhile is falling into the updraft which slows down and stops the updraft, effectively killing the convective cell.

Convection requires lifting and moving of air packets, which means that there needs to be a continual conversion of energy from potential to kinetic. It is possible to calculate an upper bound on the potential energy available for the convection of an air parcel. This is known as the Convective Available Potential Energy (CAPE) and is simply

$$\mathrm{CAPE} = \int_{\mathrm{LFC}}^{\mathrm{LNB}} \mathbf{F} dz, \tag{4}$$

where the forces acting per unit mass of the parcel ($\mathbf{F}$) are as a result of gravity and the surrounding air pressure. Equation (4) is evaluated along the path of vertical displacement, where dz is the unit scalar along this path. It is simply the work released in lifting a packet of air. This is calculated from the position with which a parcel in the boundary layer (first km above ground level) becomes buoyant, the LFC, to the LNB, where the parcel has no potential energy left. By calculating the CAPE, one can estimate whether convection can occur when an air packet reaches the LFC.

Moist convection can organise itself from single cells to various large structures. A common structure is a squall line that usually accompanies a cold front which occurs when a cool air mass follows a warmer air mass. As the cool air mass is denser, it is pushed underneath the warmer air mass, which as a result ascends. A squall line consists of lines of convective cells which are normally a few hundred km long and are created along the intersection of the warm and cold air masses. These squall lines then move with the larger air masses, creating a gust front composed of high winds at the transition region of the warm and cold air masses. As warm air is pushed upwards, this allows for condensation, resulting in cumulonimbus clouds and a region of heavy downpour behind the gust front. This convective feature can evolve over a few hours.

1.3.2 Convective weather regimes

In this thesis, convection is separated into weak and strong forcing regimes which both occur frequently in the midlatitudes, especially in the Summer. In each of these regimes, convection has either a weak or strong connection with the synoptic-scale flow. Synoptic-scale flows, which have horizontal scales on the order $\mathcal{O}(1,000\text{km})$, can affect the environment within which convection might occur and likewise, convection can affect the synoptic-scale flow depending on its vertical heating and moisture profile (Kuo and Reed, 1988).

Convection can often be categorised to be occurring in a weak or strong forcing regime depending on how it 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 CIN in the atmosphere. This Convective Inhibition (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 have the form of larger convective structures such as squall lines.

The convective adjustment timescale (Done et al., 2006) is often used to indicate which regime an area is in at any one time. 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):

$$\tau_{\mathrm{c}} = \frac{\mathrm{CAPE}}{\frac{d\mathrm{CAPE}}{dt}}. \tag{5}$$

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 qualitatively set in a model to distinguish between the weakly and strongly forced convective weather regimes.

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 could be different depending on the forcing regime. As the nature of the convergence of sampling uncertainty with ensemble size is hypothesised to depend on the underlying ensemble distribution shape, it provides the premise that one forcing situation may be more prone to sampling uncertainty than the other. As an example, the rain distribution in a weak forcing situation with precipitation would likely have a longer tail than that in a strong forcing scenario because in weak forcing, the clouds and rainfall would be more sporadically distributed in the form of single cells rather than organised systems. This larger spread in the weak forcing's rain distribution then leads to less predictability but also a smaller density in the tail of the distribution and therefore a larger sampling uncertainty in the extreme quantiles according to Equation (3) would be expected. The question is then exactly **how does convergence of sampling uncertainty with ensemble size differ between the convective weather regimes of weak and strong forcing?**

1.4 Convergence of Sampling Uncertainty on the Synoptic Scale

Due to the complexity of the atmosphere and the many scales within which weather phenomena occur, the study of the atmosphere is broken up in terms of space and time scales. The convective scale is on the order of $\mathcal{O}(10\text{km})$ and consists of the individual single convective cells. These can then organise themselves into larger structures comprising moist convection as previously described. Structures on the synoptic scale (order of $\mathcal{O}(1,000\text{km})$) include extratropical and tropical cyclones (e.g. hurricanes) and fronts with a lifetime out to two weeks.

Models of different complexities are built for different scales. A synoptic-scale model will have deep convection parameterised, where deep indicates that the height of the convective cell goes at least beyond the midtroposphere (about 8km in the midlatitudes). Convective-scale models on the other hand resolve deep convection and the dynamics of the model explicitly handle the convection. At grid-sizes of approximately 2km, there exists a "grey-zone" however, where the convection is only crudely resolved.

Sampling uncertainty convergence in forecasts has been examined primarily using data from convective-scale models. It is not clear whether the convergence of sampling uncertainty with ensemble size on the synoptic scale will have the same characteristics as it does on the convective scale. The distribution shapes will likely be a major factor in determining this. Another factor will be the availability for large enough ensemble sizes to reach the asymptotic regime. On the convective scale, one can artificially increase the effective ensemble size using the neighbourhood method. On the synoptic scale however, it is less clear whether this will be effective as the grid points are more strongly correlated and as such many more grid points would be needed to increase the effective ensemble size. As the grid points need to remain statistically similar, this may be challenging as the larger the neighbourhood region, the more chance that regions with different weather regimes and therefore different statistical properties, will be included. As such, it is possible that similar convergence behaviour as on the convective scale, also occurs in the synoptic scale. It is however, unclear, and will be investigated in this thesis.

It is interesting whether convergence of sampling uncertainty proportional to $n^{-\frac{1}{2}}$ with ensemble size also occurs in ensemble data from global NWP models with synoptic scales, as the size of an ensemble at the synoptic scale is also a pressing question. I am specifically interested in asking therefore **whether the nature of sampling uncertainty convergence with ensemble size is the same for both the convective scale as well as for the synoptic scale**.

1.5 The Key Questions

As discussed, in this thesis I am interested in answering the big question of **how to know the ensemble size required to achieve the desired accuracy in your statistic of interest**. For a thorough and complete investigation, this involves answering the three following questions:

1. How does sampling uncertainty decrease with ensemble size?
2. How does convergence of sampling uncertainty with ensemble size differ between convective weather regimes of weak and strong forcing?
3. Is the nature of how sampling uncertainty converges the same for both the convective scale and the synoptic scale?

The first step to answering these questions is to develop an ensemble large enough to measure how sampling uncertainty decreases with ensemble size, beyond that of current ensemble sizes. With a larger ensemble, it will be possible to investigate aspects of the convergence of sampling uncertainty with ensemble size that are otherwise hidden with a standard sized ensemble. To do this, an idealised model which models convection will be employed. Idealised models have been used in many studies before to understand behaviour and characteristics otherwise concealed in a more complex and computationally intensive system: for example in deeper understanding of DA (Ruckstuhl and Janjić, 2018; Petrie et al., 2017) or in explaining certain features of the atmosphere such as self-aggregation (Yang, 2021) and behaviour in specific regions of the globe (Hendricks et al., 2021). By ensuring the model is replicating the atmosphere sufficiently, which will be done in this thesis by comparing the idealised model used to full NWP models, the results will also be applicable to the larger models it is a simplified version of.

With the idealised model, the relevance of the asymptotic theory to ensemble weather prediction will first be assessed by calculating the convergence of sampling uncertainty with ensemble size for a massive idealised ensemble using a bootstrapping method. Considerations will then be made on how these findings would be relevant for an ensemble from a more complex model of a significantly smaller size. How the convergence of sampling uncertainty with ensemble size depends on the convective weather regime will additionally be assessed. Once this convergence is understood and clearly established for convective-scale data, it will be investigated whether the same results hold for synoptic-scale data.

Having an understanding of how sampling uncertainty decreases asymptotically with ensemble size for forecast statistics from forecast models of different scales can make an impact. The main reason is that by knowing how sampling uncertainty converges asymptotically, one can estimate how many ensemble members one needs to limit their sampling uncertainty to below a satisfactory level. If a forecaster is interested in the mean, they may find that they need a smaller ensemble than they would have otherwise expected and likewise, a forecaster interested in an extreme quantile can estimate exactly how many more members they require to achieve a target accuracy. This furthermore allows for a more efficient distribution of computing resources, to areas which really need it in order to advance the accuracy of weather forecasts.

Chapter 2

Asymptotic Convergence of Sampling Uncertainty

The following Chapter is adapted from the publication titled "Convergence of forecast distributions in a $100,000$-member idealised convective-scale ensemble" (Tempest et al., 2023)

2.1 Background

A major way in which forecasting can be improved is by understanding how many ensemble members an ensemble needs. This understanding can allow for greater forecasting accuracy as well as a more efficient allocation of scarce computational resources. As discussed in the Introduction, ensemble size will depend upon the statistic(s) of interest as well as the forecast variable and the level of sampling uncertainty acceptable. As such, a framework is needed to estimate the ensemble size.

Previous studies have illuminated the possibility of using the nature of how sampling uncertainty converges with ensemble size, to determine what size of ensemble is required based on the the maximum level of sampling uncertainty acceptable (Leutbecher, 2019; Craig et al., 2022). In cases where asymptotic scaling of the sampling uncertainty proportional to $n^{-\frac{1}{2}}$ would be observed, it could 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 is wanting to approximate the number of members required to reach a certain accuracy in the spread of their measurement of temperature over Munich. This would work by quantifying how their data which they have with their current sized ensemble scales with $n^{-\frac{1}{2}}$, and then extrapolating this until it reaches the level of sampling uncertainty they would require. The corresponding ensemble size would then be their desired ensemble size for that level of uncertainty.

This chapter assesses the relevance of the asymptotic theory to ensemble weather prediction in it's ability to estimate what ensemble size is needed. A computationally efficient idealised 100, 000-member ensemble forecasting system is considered to build on the results from our 1, 000-member convective-scale ensemble using a full NWP (Craig et al., 2021). This idealised ensemble is checked to be realistic in terms of space and time scales, replicating convective processes as well as the shape of the forecast distributions. The ensemble sizes required to obtain the asymptotic scaling for different quantities and their dependence on the underlying distribution will be investigated as well as considering how to obtain information about convergence from ensembles of a smaller, operational, size.

In Section 2.2, 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 which are carried out on the subsequent ensemble data. Section 2.3 begins with evaluating whether the distributions from the idealised prediction system are of a similar shape as those from a full NWP model. The results of exploring the convergence behaviour are then reported in Section 2.3.2. In Section 2.3.3 two methods are introduced which determine whether one is scaling asymptotically as well as how to estimate the sampling uncertainty convergence at larger ensemble sizes using only a smaller ensemble. The main results are then summarised in Section 2.4.

2.2 Model and Methods

A model is required which represents the basic processes of convection in the midlatitude atmosphere. This encompasses having space and time scales representative of convective processes and being capable of modelling non-linear processes. It must, in addition, be computationally inexpensive, so that ensemble sizes of order $\mathcal{O}(10^5)$ can be examined efficiently. A one-dimensional idealised model for cumulus convection (Wuersch and Craig, 2014) is employed, 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 non-linear dynamics of the convective life cycle and the spatially intermittent and non-Gaussian statistics of a convecting atmosphere. In Section 2.2.1 the model of (Wuersch and Craig, 2014) is described, and in Section 2.2.2 it is assessed whether this model achieves the requirements stated above. The idealised prediction system built on the basis of the idealised model is presented in Section 2.2.3, before the methods used are outlined in Section 2.2.4.

2.2.1 Idealised model

The one-dimensional idealised model (Wuersch and Craig, 2014) 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

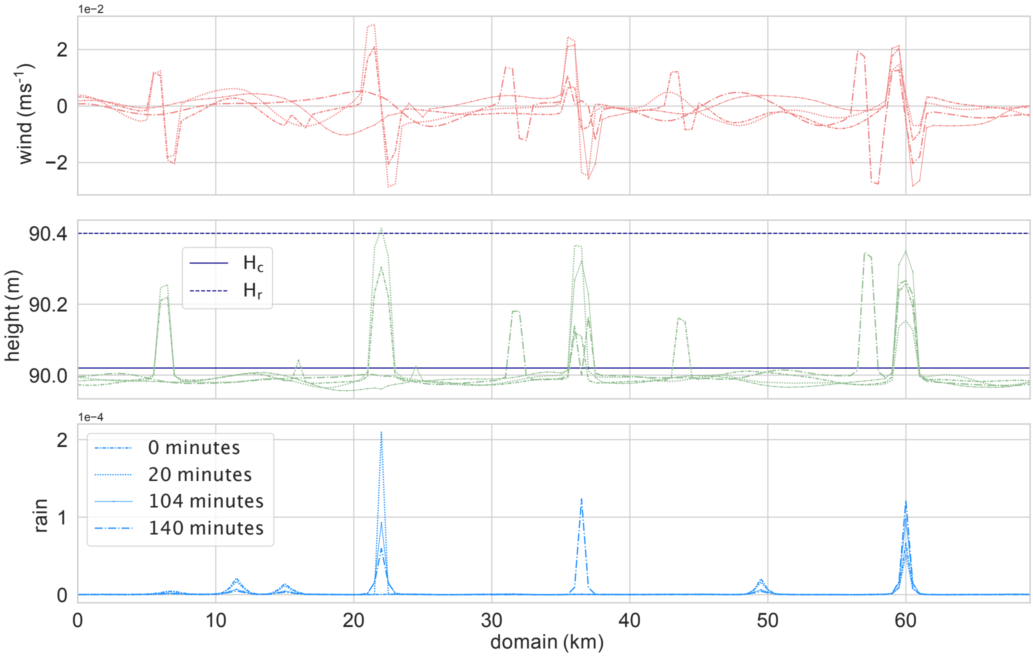

Figure 2.1: Snapshots of the domain at four time points for the three model variables. Thresholds (described in text) are shown in the height field.

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 2.1. These are described by the following equations:

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + \frac{\partial(\phi + c^2 r)}{\partial x} = K_u \frac{\partial^2 u}{\partial x^2} + F \tag{4}$$

$$Z = h + H \tag{6}$$

$$\phi = \left\{ \begin{array}{ll} \phi_c + gH, & Z > H_c \\ g(H+h), & \text{otherwise} \end{array} \right\} \tag{7}$$

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = K_h \frac{\partial^2 h}{\partial x^2} \tag{8}$$

$$\frac{\partial r}{\partial t} + u\frac{\partial r}{\partial x} = K_r \frac{\partial^2 r}{\partial x^2} - \alpha r - \left\{ \begin{array}{ll} \beta\frac{\partial u}{\partial x}, & Z > H_r \text{ and } \frac{\partial u}{\partial x} < 0 \\ 0, & \text{otherwise} \end{array} \right. \tag{9}$$

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 orography is not included, so that

$H = 0$ and therefore $Z = h$. From selecting the initial fluid level height, h_0, to be 90m, the gravity wave speed is 30ms^{-1}, as in Wuersch and Craig (2014). The diffusion constants used are: $K_u = 2 \cdot 10^3\text{m}^2\text{s}^{-1}$, $K_h = 6 \cdot 10^3\text{m}^2\text{s}^{-1}$ and $K_r = 10\text{m}^2\text{s}^{-1}$.

If h is greater than a first threshold ($h > H_c = 90.02\text{m}$), 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 $899.77\text{m}^2\text{s}^{-2}$. This encourages more fluid into this region, thereby increasing h further. This process is analogous to the developing, buoyant updraft phase of a cloud whereby the LFC 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 ($h > H_r = 90.4\text{m}$), and wind is converging on this grid point, then rain (scaled by β which is set to 0.1) is produced. This is the mature stage. 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 and the dissipation stage. Rain is removed from the domain by a linear relaxation of rate α, with value $1.4 \cdot 10^{-4}\text{s}^{-1}$. 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 $x = 22\text{km}$ in Figure 2.1. The height crosses the rain threshold at $t = 20$ minutes, the negative buoyancy due to the rain changes the convergent wind to divergent, and the height perturbation has disappeared by $t = 104$ minutes 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 1^{st} order derivative of a Gaussian function. This odd function is multiplied by an amplitude, $\bar{u}$, which has value $8.95 \cdot 10^{-3}\text{ms}^{-1}$. 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 Wuersch and Craig (2014), 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 4s 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.

2.2.2 Properties of the model solutions

The example in Figure 2.1 show that the evolution of the simulated cloud life cycle occurs on realistic time scales. For the updraft phase of a cloud, the time between a cloud's

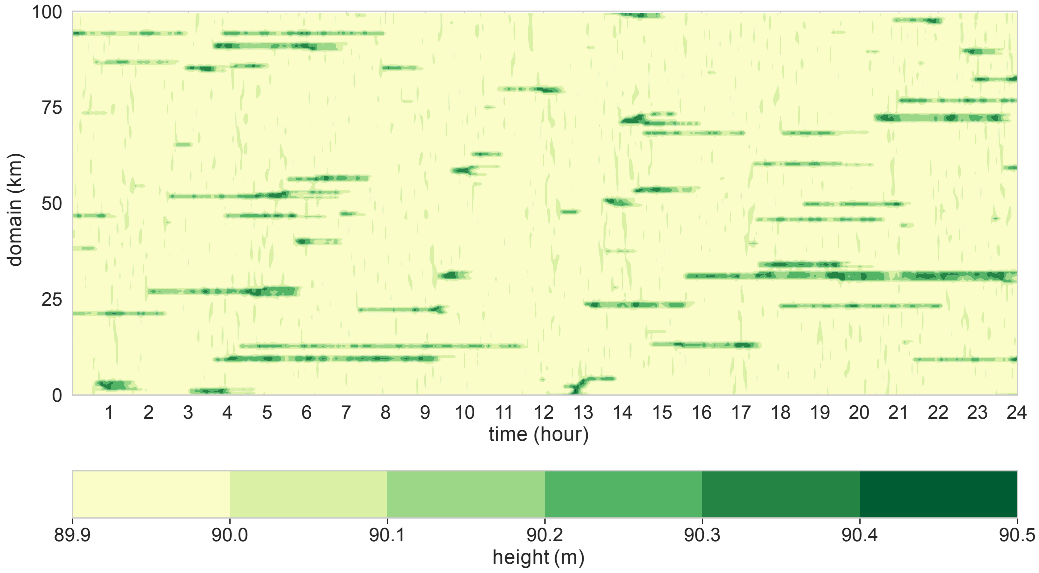

Figure 2.2: Hovmöller diagram showing the evolution of the height field across a section of the domain over the 24 hour period of the free forecast.

initiation ($h > H_c$) and rain formation ($h > H_r$) is approximately 15 minutes. For the downdraft phase, the half-life of rain is approximately 1 hour and the overall lifetime of a cloud ($h > H_c$) with one updraft and one downdraft, is between 1 and 2 hours. Multiple phases of a cloud can exist, as shown in Figure 2.2 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 500km and a horizontal resolution of 500m, 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 1.2km and a maximum width of 7.5km, which is in agreement with Wuersch and Craig (2014). 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 to the theory detailed in Craig and Cohen (2006) and 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.

2.2.3 Idealised prediction system

An 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 90m, and all simulations are run for $1,000$ time steps with independent realisations of the stochastic forcing term to spin up (the short initial adjustment period) the model fields.

After initialisation, the ensemble Kalman filter (EnKF) DA (Evensen, 1994) is cycled 50 times. Observations were assimilated every 5 minutes 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 the localisation radius as 2km. For more details on the DA used in this system, see Ruckstuhl and Janjić (2018) and Ruckstuhl et al. (2021). After 50 cycles the 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 set-up 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 which 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 hours, 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 idealised prediction system described here models different sources of forecast error. The EnKF provides 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.

2.2.4 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.

Non-Gaussian statistics

To test how close the forecast distributions are to being Gaussian distributed, the same measures as used by Kondo and Miyoshi (2019) are employed. These are sample skewness, sample excess kurtosis and the Kullback-Leibler Divergence (KL Divergence) (Kullback and Leibler, 1951). 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 non-symmetric measure of the difference between two distributions and is used to measure the distance a histogram of a distribution from the ensemble is, from that of a reference Gaussian Probability Density Function (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).

Statistical inference

Each finite-sized data set $(x_1, x_2, ..., x_n)$ of length n, created by an ensemble with n members is just one realisation of the random variables $(X_1, X_2, ..., X_n)$ from a distribution F, and, as such, each of the sample statistics (e.g. sample mean $\bar{x}_n = \frac{x_1+x_2+...+x_n}{n}$) is just one possible realisation of a random variable (e.g. $\bar{X}_n = \frac{X_1+X_2+...+X_n}{n}$) (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 $\bar{X}_n$) will determine the associated uncertainty of the estimation.

If this underlying distribution F is unknown, non-parametric bootstrapping (Davison and Hinkley, 1997) is a powerful tool used to infer information about its characteristics. Bootstrapping assumes that the estimate $\hat{F}$ is an accurate realisation of F. Non-parametric bootstrapping is re-sampling with replacement from a data set where all data points have equal probabilities $\frac{1}{n}$, to create a "bootstrapped" random sample $(X_1^*, X_2^*, ..., X_n^*)$, 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 $\bar{x}_n^*$). The distribution of this statistic (the random variable of the bootstrapped mean $\bar{X}_n^*$) can then be used to construct CIs, 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 $\bar{X}_n$, is known as the bootstrap principle (Dekking et al., 2005).

For the analysis of uncertainty, 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 CI, the percentile method is employed where for the 95% level, the 2.5$^{\text{th}}$ and 97.5$^{\text{th}}$ 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. The width of the 95% CI is then labelled the "convergence measure".

2.3 Results and Discussion

2.3.1 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 evolutions and shapes analysed and compared with distributions extracted from our $1,000$-member full NWP ensemble (Craig et al., 2021).

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 subsections therefore will show distributions of the three variables of the idealised model for both initially cloudy and noncloudy grid points.

Evolution of the wind variable

Figure 2.3 shows distributions of the wind variable at four time points in the evolution of the free-run at an initially cloudy, and noncloudy, grid point. In each histogram, 100 bins are calculated in order to clearly resolve the shapes of the distributions. The histograms

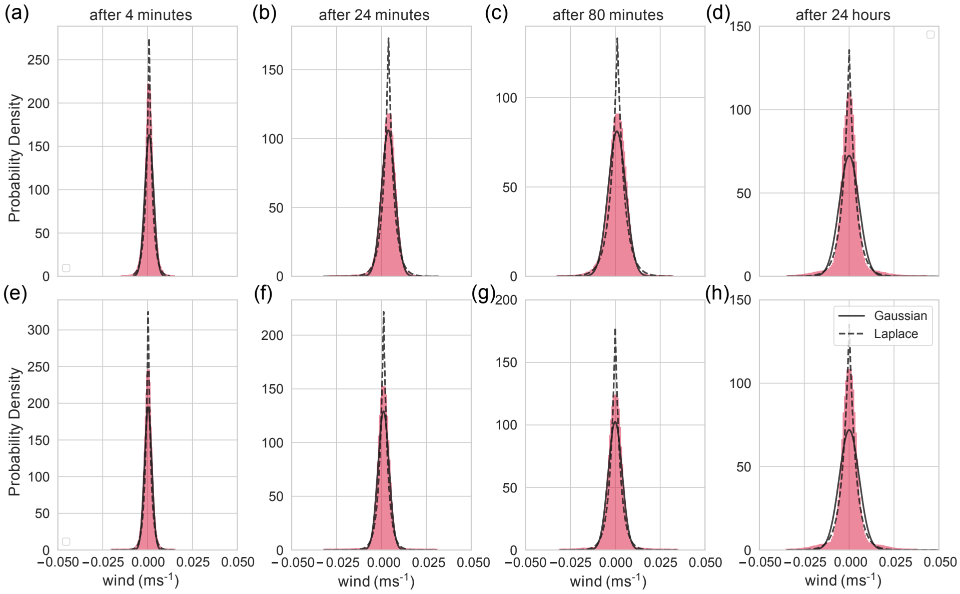

Figure 2.3: Wind variable distributions from the 100, 000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 minutes, and (d,h) at 24 hours of free run, overlaid by a Gaussian and Laplace PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 2.1

Starting conditions	After: 4 minutes	24 minutes	80 minutes	24 hours
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]

Table 2.1: Non-Gaussian statistics of wind distributions. Entries of table are [skewness, kurtosis, KL Divergence].

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 which is extracted from an initially cloudy grid point shows an increase in spread and tail density until 80 minutes. At 24 minutes there is a shift in the mean towards positive wind, but the mean relaxes gradually to zero again as seen at 80 minutes. The distribution at 24 hours is centred around zero. At the initially noncloudy grid point, the distribution follows a similar evolution, except at 24 and 80 minutes where the mean remains near zero. Table 2.1 documents three statistics that characterise the non-Gaussianity of the distributions presented in Figure 2.3. It is clear from the kurtosis that density increases at the tails and the distributions at both grid points become slightly less Gaussian as time evolves. It is interesting to note that the kurtosis of the distribution at the grid-point which began with no cloud increases at a faster rate than at the grid point which started the free-run with a cloud. Also clear is the symmetry of the distributions (small skewness) throughout the evolution.

At all time points and for both grid points, KL Divergence (Table 2.1) is below 0.3 and as such a Gaussian PDF fits well to the distributions. Figure 2.3 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 minutes 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.

Evolution of the height variable

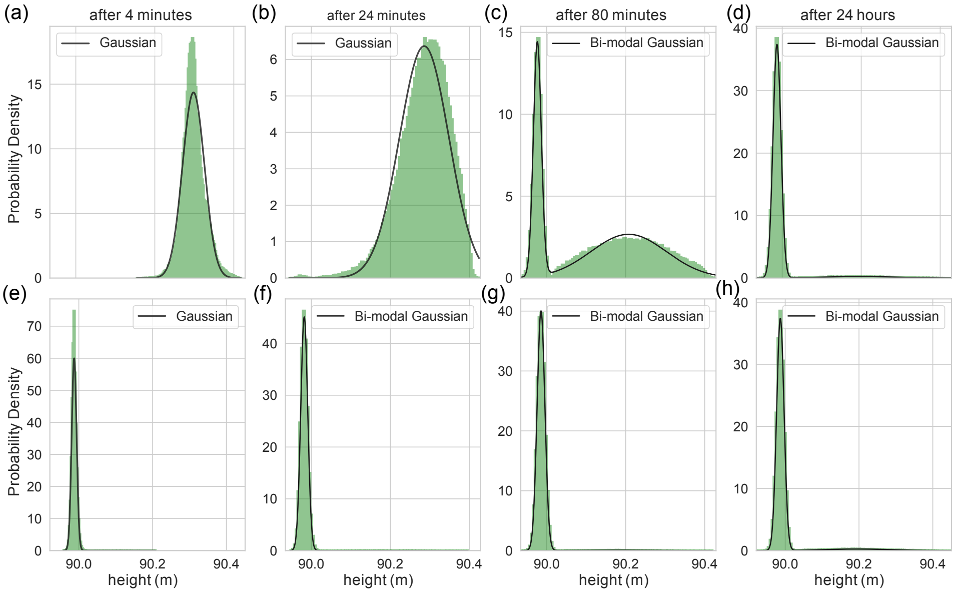

Figure 2.4: Height variable distributions from the 100, 000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 minutes, and (d,h) at 24 hours of free run, overlaid by a Gaussian or bimodal Gaussian PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 2.2

Starting conditions	After: 4 minutes	24 minutes	80 minutes	24 hours
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]

Table 2.2: Non-Gaussian statistics of height distributions. Entries of table are [skewness, kurtosis, KL Divergence].

As with the wind variable, the evolving shapes of the height variable distributions (Figure 2.4) are analysed. The histogram of the height variable at 4 minutes at the grid point initially containing a cloud shows a single peak with mean above the H_c threshold of 90.02m. As the ensemble 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 minutes, but is clearly visible by 80 minutes. The formation of a second peak is accompanied by a large increase in KL Divergence (Table 2.2). 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 to the noncloudy peak and the bi-modality is hardly visible. The evolution of the distribution for the first 80 minutes at the grid point which 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 minutes, the distributions at both grid points still show the approximately Gaussian distribution produced by the DA. After 24 and 80 minutes, a bi-modal Gaussian fits well to the distributions, for both the grid points which started without and with a cloud. This deviation from a simple Gaussian is consistent with the increase in KL Divergence to be above 0.3 (Table 2.2) at these grid and time points.

Evolution of the rain variable

Starting conditions	After: 4 minutes	24 minutes	80 minutes	24 hours
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]

Table 2.3: Non-Gaussian statistics of rain distributions. Entries of table are [skewness, kurtosis, KL Divergence].

The evolution of the rain variable distributions are shown in Figure 2.5. Note the log scaled x-axis. The ensemble members below a certain threshold ($3 \cdot 10^{-5}$) 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 clearly observe this reduced number of members. In the case of the grid point beginning with a cloud, it contained a cloud which had not yet precipitated. At 4 minutes therefore, many of the ensemble members had not yet precipitated. The fraction of members with rain increases up until 80 minutes, at which time 75% of the members contain rain, compared to 4% at 4 minutes. Rain is removed by a sink (decreasing) function that is proportional to the rain amount, so that the largest rain amounts experience the most rapid decline, with the results that the peak of the distribution is shifting towards smaller values between 24 and 80 minutes. This is also seen in the strong increase of skew in Table 2.3. As the members at 24 hours become decorrelated, there is no characteristic cloud size which would have

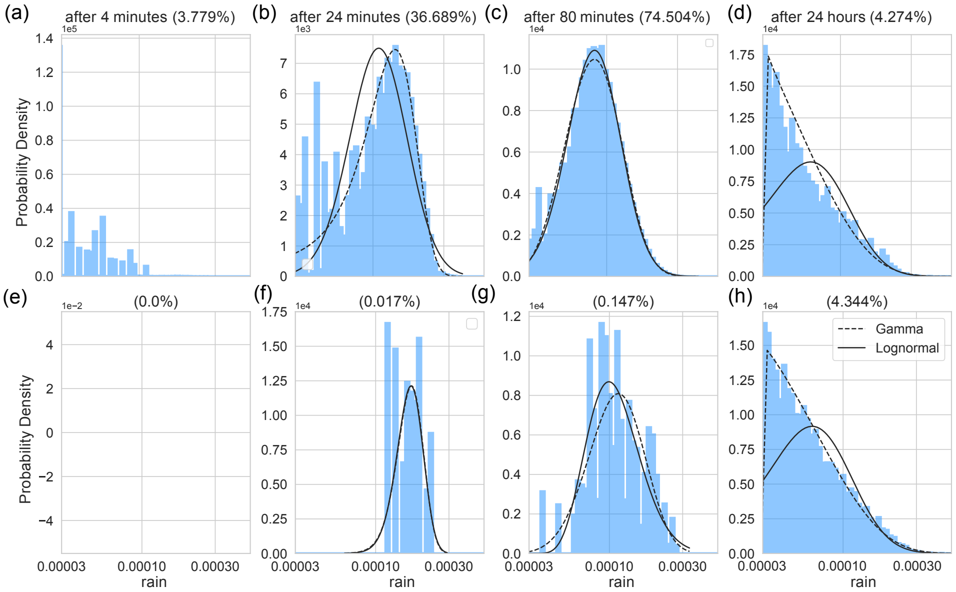

Figure 2.5: Rain variable distributions from the 100,000-member ensemble at initially (a,b,c,d) cloudy and (e,f,g,h) noncloudy grid points after (a,e) 4, (b,f) 24, and (c,g) 80 minutes, and (d,h) at 24 hours of free run. Percentages above histograms show the number of members containing rain, which is shown in the histogram. Overlaid by gamma and lognormal PDF. Non-Gaussian statistics corresponding to the distributions are detailed in Table 2.3.

resulted in a well-defined peak as seen at 24 and 80 minutes. A similar evolution occurs at the grid point which 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 minutes. 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 2.3.

When there is significant rain ($> 0.1\%$ of members), the rain distribution fits well to a Gamma PDF, and to a lesser extent, a log-normal PDF. 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 2.5(f) contains only 17 members with rain, however it appears to also be able to be approximated by a Gamma/Log-Normal PDF.

Comparison with NWP models

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. h can therefore correspond to the saturation deficit, or relative humidity, variables that capture the atmospheric variability inside and outside of 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 our study of a $1,000$-member ensemble forecast using a NWP model Craig et al. (2021), where it was found that the distributions of all the examined forecast variables fell into one of three broad categories: quasi-Gaussian, multi-modal or highly skewed. The parameterised fits for the three variables of the idealised model are thus representative of the three categories that characterise NWP ensemble forecasts. Furthermore, the evolution in time of the model variable distributions follows our conceptual model as described in the Introduction. Based on these results, it is anticipated that the convergence characteristics of the distributions will also be representative of the behaviour of real-world NWP systems.

A preliminary analysis of the bivariate distributions was carried out in addition (not shown). 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 the wind at two different time points where similar structures were created to those from the idealised model employed by Poterjoy (2022).

2.3.2 Sampling uncertainty convergence

The convergence of sampling uncertainty of statistical properties as ensemble size increases is now analysed. Following our previous study with the $1,000$-member NWP ensemble, statistical inference is carried out on selected uni-modal distributions from the ensemble in the free-run component of the idealised prediction system to identify the nature of the convergence of sampling uncertainty. It is further investigated how sampling uncertainty convergence is sensitive to the shape of the distribution and the statistic being evaluated.

Universal convergence scaling characteristic

The analysis of convergence will focus on two cases: the 80 minute forecast for an initially cloudy grid point, and the 24 hour forecast for an initially noncloudy grid point. As can be seen from panels (c) and (h) of Figures 2.3, 2.4 and 2.5, these two cases include the three main distribution types found in the forecasts. Note that for the rain distributions, the zero-rain data points that are omitted from the distribution plots are included in all convergence measure 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 to 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 until 100, 000 members, the random variable sampling distribution is calculated in steps of 100 members. The width of the CI (between the 2.5th and 97.5th percentile of the random variable sampling distribution), which is defined 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 $y = an^{-\frac{1}{2}}$ using linear regression in log space, where a quantifies how the convergence measure scales with n. The range of values used for each fit are detailed in the Appendix (Section 6.1). A forecast statistic is described as being in the asymptotic regime if the convergence measure appears to be converging as $n^{-\frac{1}{2}}$ with ensemble size.

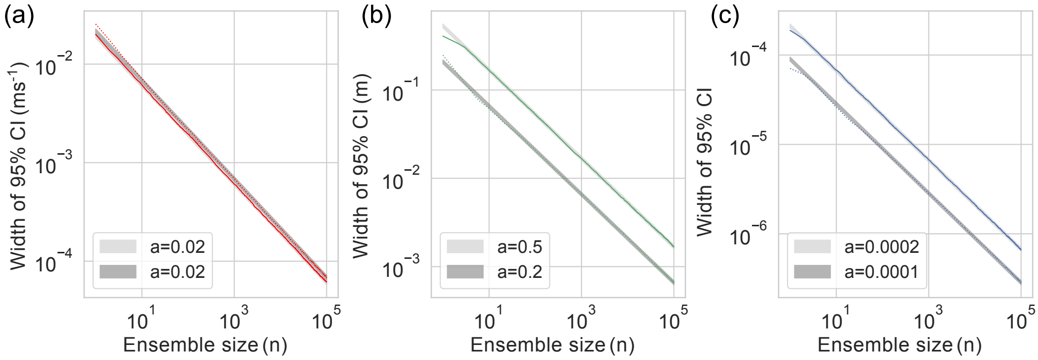

Figure 2.6: Continuous and dotted coloured lines are width of 95% CI of the sampling distribution of the mean for (a) wind, (b) height and (c) rain model variables. The continuous line uses distributions from (c) of Figures 2.3, 2.4 and 2.5. The dotted line uses distributions from (h) of Figures 2.3, 2.4 and 2.5. Light and dark grey lines are fitted to continuous and dotted lines respectively, see the text for details. The corresponding width of grey lines spans 5% above and below fitted line. The fitted parameter is shown in the legend. The number of ensemble members used for fitting are catalogued in the Appendix, Section 6.1.

The convergence measure of the mean, as a function of ensemble size, n, is shown in Figure 2.6 for the three model variables for the two cases. The fitted power law lines which scale as $n^{-\frac{1}{2}}$ follow the width of the 95% CI well for each distribution and model variable, except at ensemble sizes below 10 for the height and rain distributions. The decrease of the sampling uncertainty of the sample mean proportional to $n^{-\frac{1}{2}}$ 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 CI is not. Finally, it is noted that the convergence measures are similar for both rain distributions which included, and did not include, the 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.

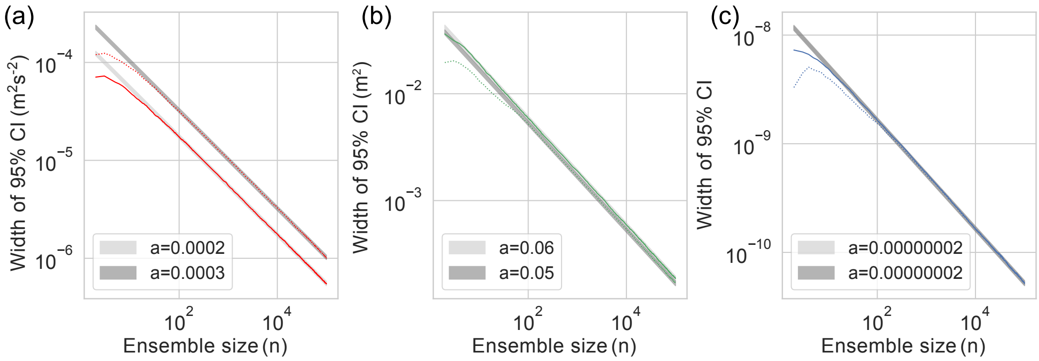

Figure 2.7: As in Figure 2.6 but for the sampling distribution of the variance.

The convergence measure for the variance is shown in Figure 2.7. The power law scaling of $n^{-\frac{1}{2}}$ is seen again in all distributions. As expected, the CLT is not only applicable to the mean, but also to other forecast statistics. The number of members required until convergence appears to follow $n^{-\frac{1}{2}}$ is generally larger than for the mean (Figure 2.6), and there is an overestimation of the width made by the fit at smaller ensemble sizes. This is in line with our previous study with the 1,000-member NWP ensemble where we discovered that more members are required in the standard deviation compared to the mean in order to achieve convergence as predicted in the asymptotic limit.

The convergence of various quantile, p, sampling distributions are shown in Figure 2.8. With enough members it is clear that in most cases the convergence measure scales as $n^{-\frac{1}{2}}$, with wider CIs 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 $n^{-\frac{1}{2}}$ scaling of sampling uncertainty with ensemble size as long as enough members are used. The exception was

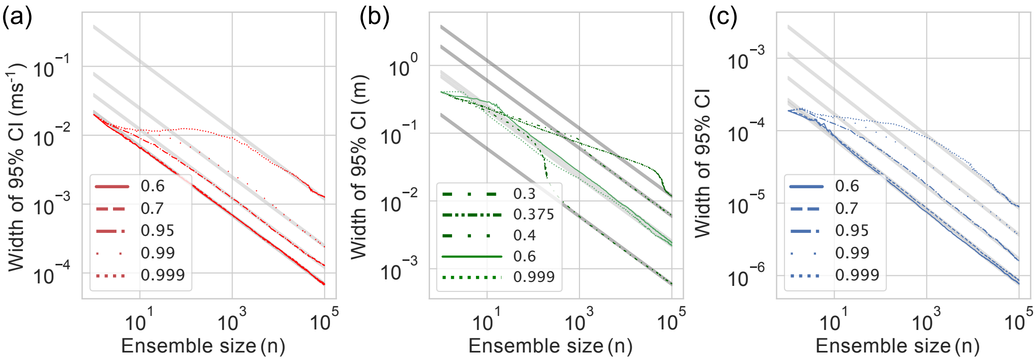

Figure 2.8: As in Figure 2.6 but for sampling distributions of different quantile levels, p. Legend labels the different quantiles.

the 0.999 quantile. It could be seen to scale approximately as $n^{-\frac{1}{2}}$ 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. It will be seen 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 distribution shape in this region.

Dependence on distribution shape

It has been seen that the convergence measure scales proportional to $n^{-\frac{1}{2}}$ with ensemble size for a sufficiently large ensemble. However, the constant a, and hence the absolute width of the CI, depended on the forecast statistic and on the case being considered. To better understand these results, this section will systematically investigate 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 2.3 (c) and (h)). The width of the CIs for the estimates of the means are also very similar (Figure 2.6(a)). When the distribution shapes are less similar, as for the height and rain distributions in Figures 2.4 and 2.5, 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 multi-modal or skewed distributions of the other variables have larger density away from the mean.

The width of the 95% CI for estimates of the ensemble variance also shows differences between the two cases (Figure 2.7), 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 again may relate to where the density of the underlying distribution is located, but the connection is less clear.

For convergence measures of the quantile estimates shown in Figure 2.8, the majority of convergence lines are offset from one another. For the unimodal distribution of the wind and rain, the further the quantile is from the median, the larger the width of the 95% CI. Hence more uncertainty is attached to these quantiles at the tails compared to at the centre of the distribution. This behaviour is expected from Equation (3), which states that the standard error of a quantile estimate will be inversely proportional to the density of the underlying distribution at that quantile level. The behaviour for the height variable is more complex, with large sampling uncertainties for intermediate quantile levels. This is also consistent with Equation (3) however, since the bimodal distribution of height has a minimum near the p=0.375 quantile, leading to wide CIs 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 (standard error) for a range of quantiles from a distribution. The values are computed using Equation (3), 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 (Scott, 2015). 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 (3) 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 which are calculated using the bootstrap method employed previously, with three subensemble sizes (50, 100 and 500 members).

Figure 2.9 shows the resulting contour plot for the near-Gaussian wind distribution (Figure 2.3(c)), which resembles the result for a true Gaussian in Figure 1.4. 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 members required to reach smaller levels of uncertainty increases. As expected, the tails of the distribution are more uncertain compared to the peak of the distribution in this uni-modal 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 2.8(a), particularly observable at the 0.95 and

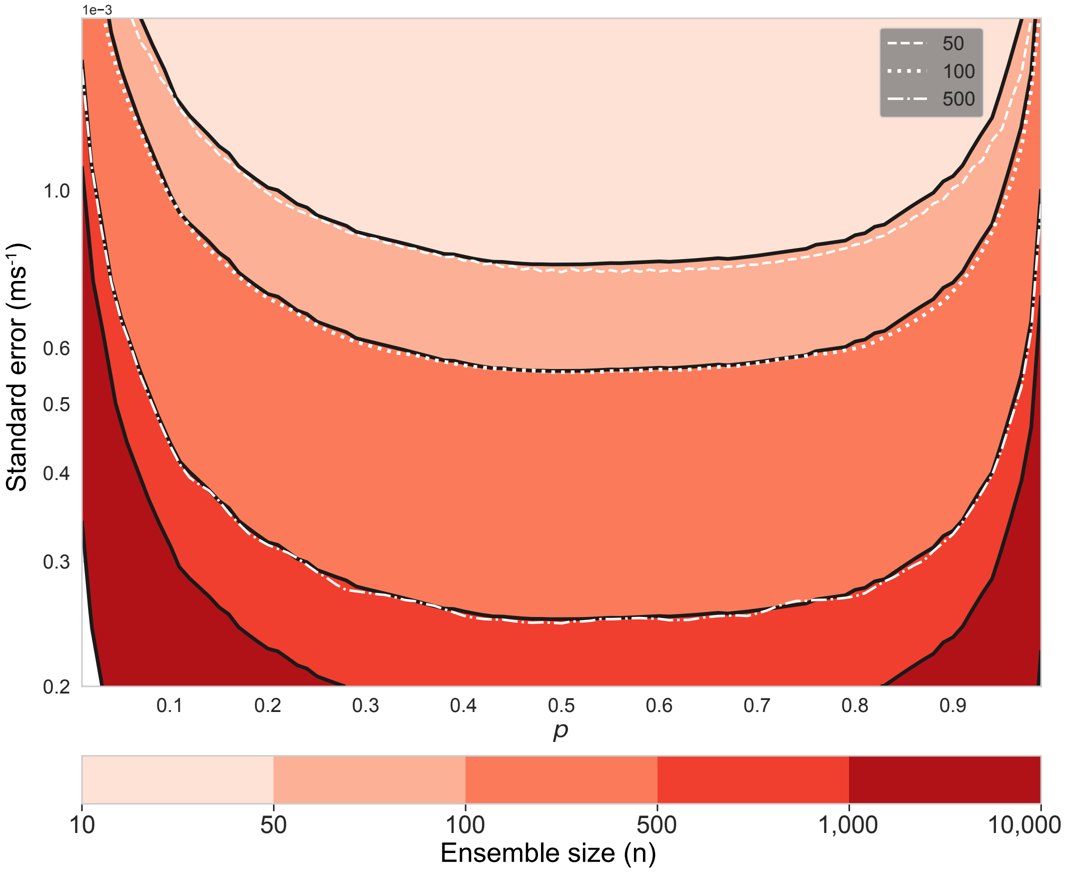

Figure 2.9: Contours show the number of members required to achieve a standard error (y-axis) for quantile levels ranging from 0.01 to 0.99 in steps of 0.01 (x-axis) for the distribution of Figure 2.3(c). White lines show an estimate using the bootstrapping technique.

0.99 quantile levels. As the uncertainty calculated in Equation (3) is proportional to $n^{-\frac{1}{2}}$, large deviations between the contours and white lines indicate that the bootstrapped data is not yet converging as $n^{-\frac{1}{2}}$ for that given ensemble size.

As with the wind distribution, a contour plot of n is calculated for the height distribution (Figure 2.4(c)) and is visualised in Figure 2.10. 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 2.10 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 (9). Any grid points that surpass this height immediately

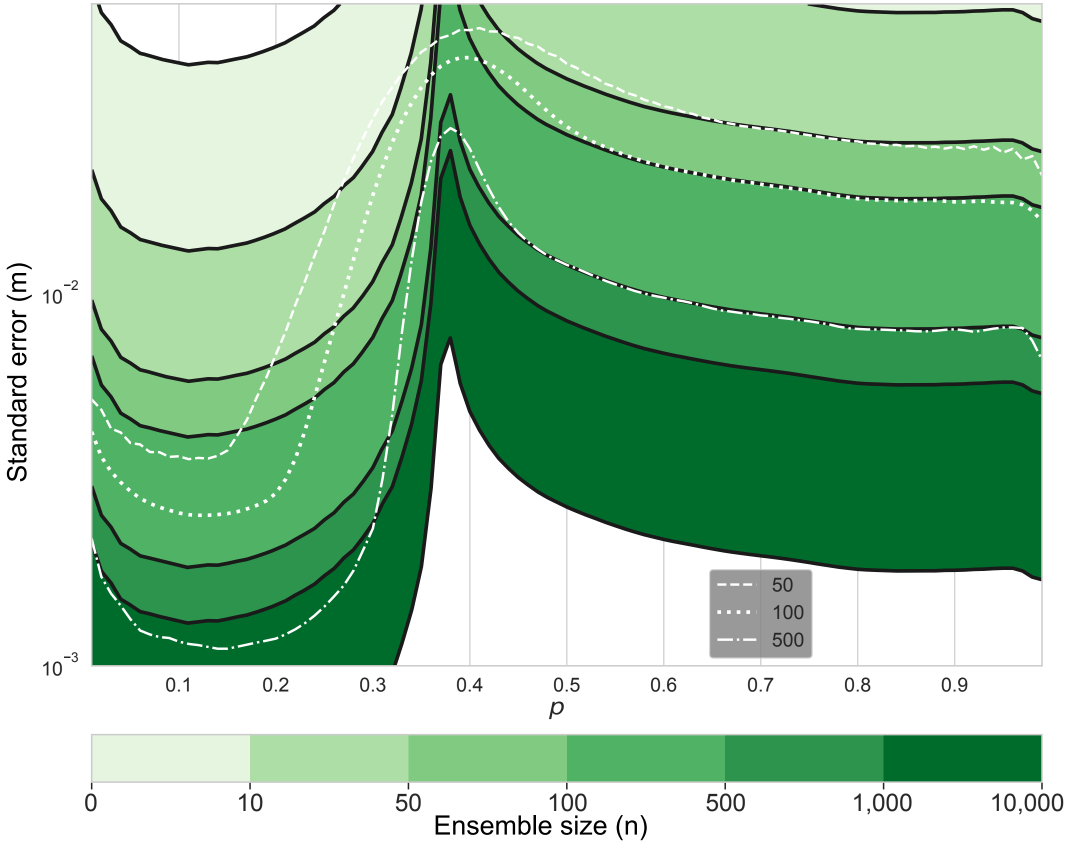

Figure 2.10: As in Figure 2.9 but with a height distribution from Figure 2.4(c).

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 2.5(c)) is shown in Figure 2.11. 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 2.10, 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.

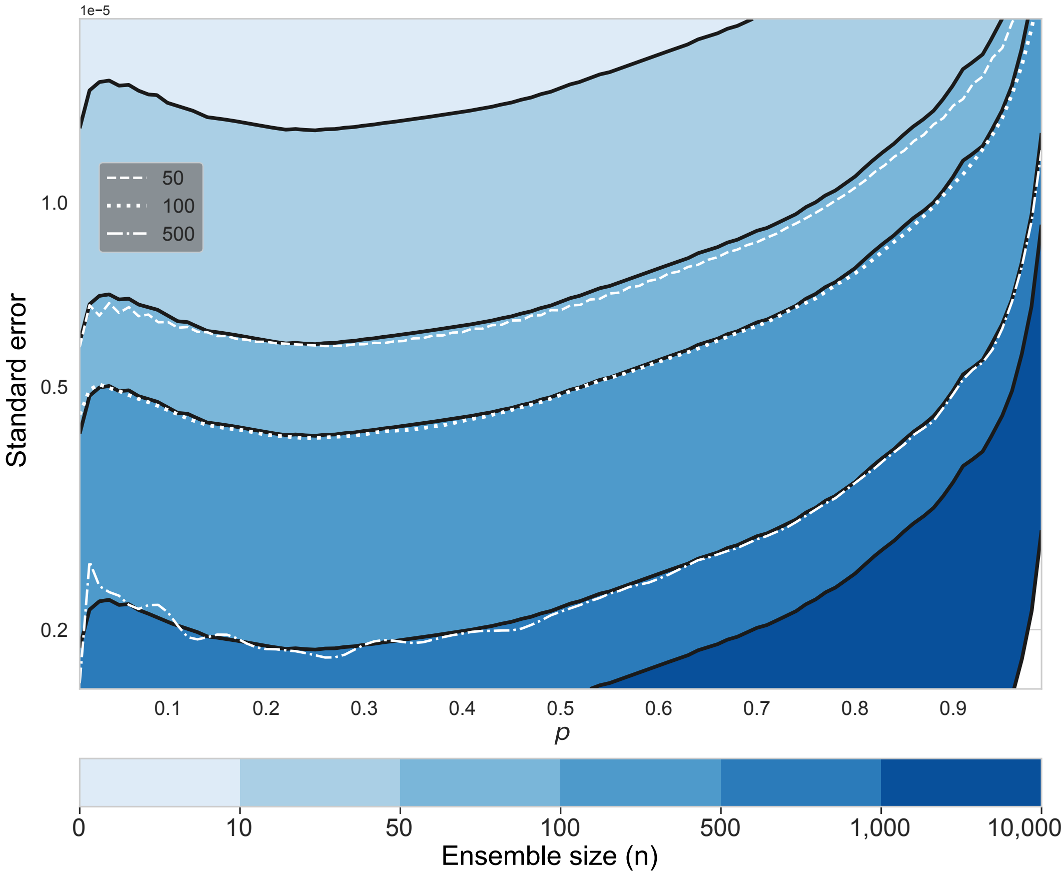

Figure 2.11: As in Figure 2.9 but with a rain distribution from Figure 2.5(c).

2.3.3 How big an ensemble do I need?

An important benefit of the asymptotic scaling of the width of the CIs is that an estimation can be made of the number of ensemble members needed to reduce sampling uncertainty to a desired level. 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, two approaches to estimating convergence properties when only small ensembles are available, will be explored. First, it is considered 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 precisely computed.

Bootstrapping using smaller ensemble sizes

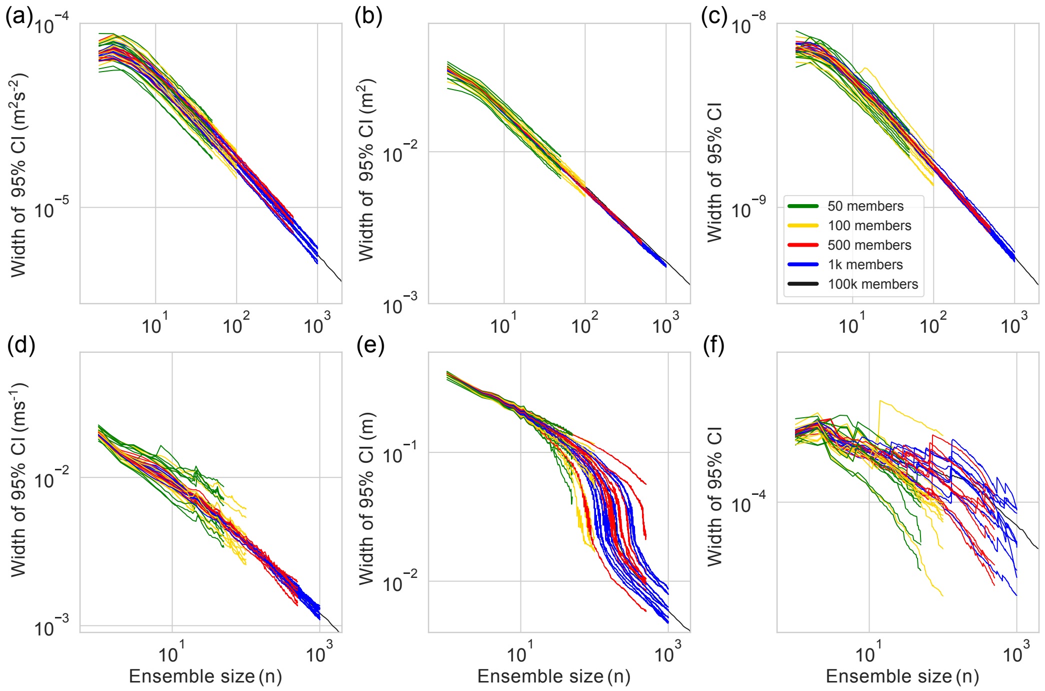

Figure 2.12: Width of 95% CI of sampling distribution of (a) variance and (d) 95th percentile of wind distribution (Figure 2.3(c)), (b) variance and (e) 30th percentile of height distribution (Figure 2.4(c)) and (c) variance and (f) 99.9th percentile of rain distribution (Figure 2.5(c)) as a function 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-member distribution. The convergence measure calculated using all 100,000 members is in black in the background.

If only a small ensemble is available for a forecast, it is still possible to construct a bootstrap estimate of CIs as before, but these estimates may not be useful if the small ensemble is not representative of the full distribution. To investigate this issue, CIs are first constructed which are based on different small ensembles drawn from the 100,000 members computed previously, to see whether the convergence behaviour is consistent. Figure 2.12 shows convergence curves for a sample of forecast variables, namely the variance and selected quantiles of the wind, height and rain distributions (see Figures 2.3(c), 2.4(c) and 2.5(c) respectively). This includes variables that converge for relatively small ensemble sizes, as well as more extreme values that occur only rarely. The plots show the convergence measure 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 convergence measures constructed from the 100,000-member ensemble are also plotted.

For the variables on the top row of Figure 2.12, even 50 members is sufficient to identify the asymptotic convergence regime with the width of the CI scaling as $n^{-\frac{1}{2}}$. It is interesting that the correct scaling behaviour is found for the estimates based on the smaller ensemble sizes, although there is spread in the constant offset of the curves that increases as the ensemble becomes smaller. Figure 2.12(d) shows an example where the 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 $n^{-\frac{1}{2}}$ slope by chance. Figure 2.12(e) 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 the asymptotic behaviour for a limited range of n before dropping to the true convergence measure 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 2.12(f) shows no evidence of convergence for any of the ensemble sizes considered here.

Figure 2.12 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 CI, 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 the question is posed of whether a set of ensembles generated this way show the same variability as an ensemble drawn from the full distribution.

Figure 2.13 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 if a given ensemble forecast is large enough to produce the asymptotic scaling behaviour. If this is the case, the estimate of the sampling uncertainty can be reliably extrapolated to predict how sampling uncertainty will decrease with ensemble size.

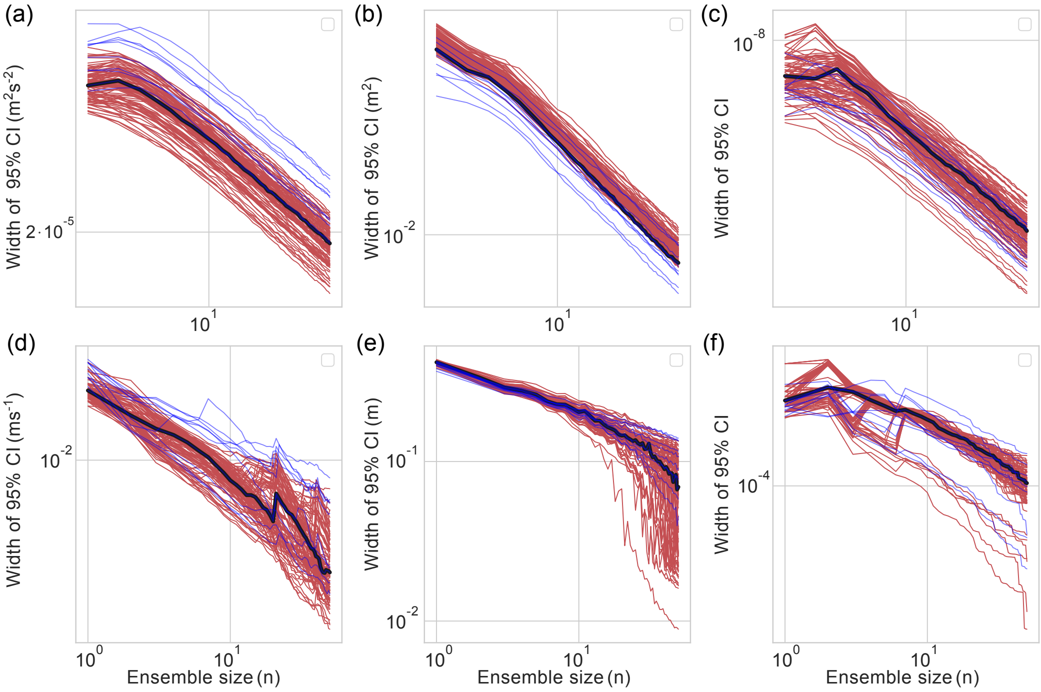

Figure 2.13: (a-f) as in Figure 2.12. Blue lines are green lines from Figure 2.12. One sampled distribution of size 50 from the original 100, 000-member ensemble was bootstrapped to obtain 100 distribution samples of size 50. The convergence measure calculated from these distributions is shown in red. The sampled distribution used for red lines has its own convergence measure shown in black.

Parameterisation of distributions

As it has been 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. But for many quantities of interest, 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 the potential of using a priori knowledge of the distribution of a forecast variable is explored to provide improved estimates from such small ensembles. Figures 2.3, 2.4 and 2.5 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 like 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, whose resulting convergence will be compared. From the results, it can be concluded whether the parameterisation technique can calculate the convergence measure accurately, and how accurate it is when only 50 members are used.

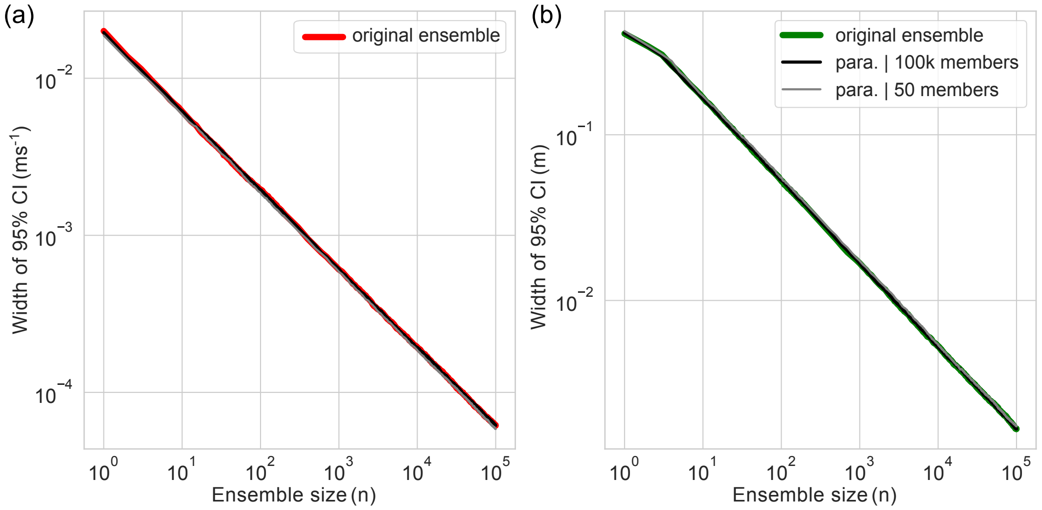

Figure 2.14: (a) Red and (b) green lines are width of 95% CI of sampling distribution of mean for (a) wind distribution (Figure 2.3(c)) and (b) height distribution (Figure 2.4(c)), calculated using bootstrapping using the 100, 000-member ensemble data. Black lines show convergence using data generated 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.

The convergence measure of the mean calculated from using distributions generated from a parametric fit of a wind and height distribution (Figures 2.3(c) and 2.4(c), respectively) is shown in Figure 2.14. The parameterisations (Gaussian for wind and bi-modal Gaussian for height) which used two different sizes of ensemble for parameterisation (each shown in grey and black lines) showed good agreement to the convergence calculated using the original 100, 000-member ensemble data.

The convergence measure of the variance could be approximately estimated by parameterisation of the ensemble's distribution (Figure 2.15). 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 which has smaller variance. This leads to the resulting sampling distribution of variance to be smaller and hence produce a narrower 95% CI. This is also the reason for the shifting in the case of the height distribution. Note that using more than 50 members for the parameterisation does not improve the results greatly.

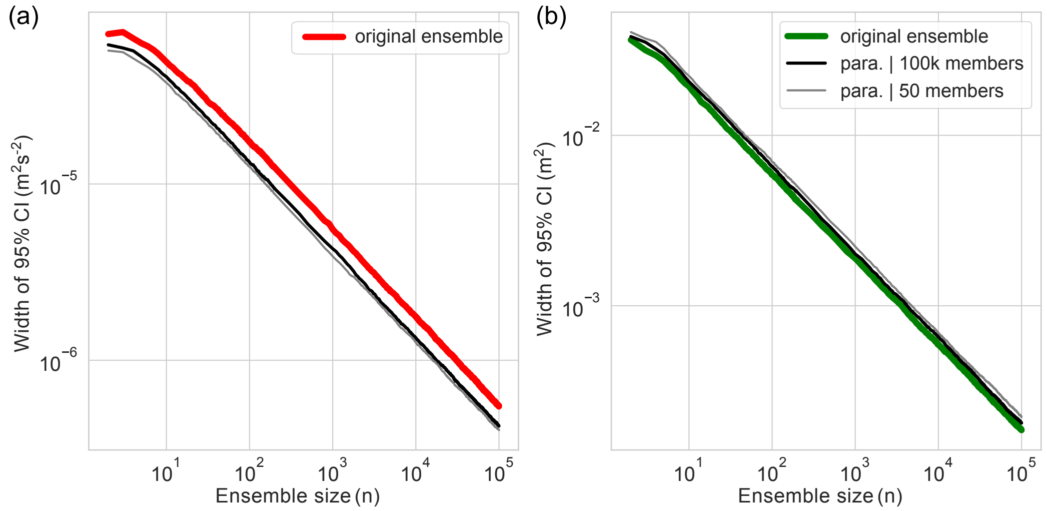

Figure 2.15: As in Figure 2.14, but for the sampling distribution of the variance.

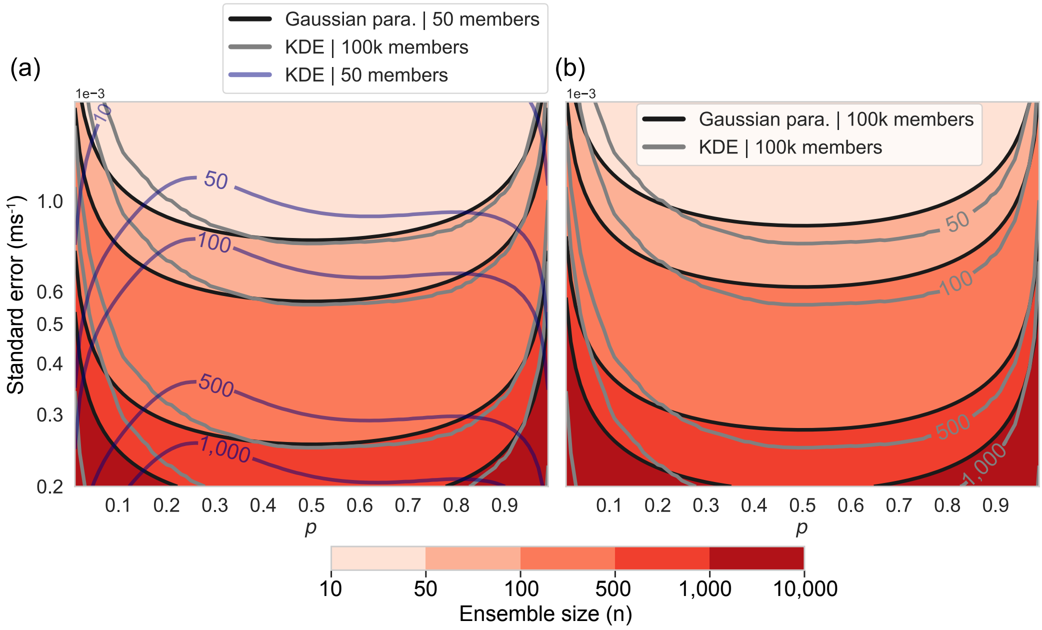

Figure 2.16: Black contours created as in Figure 2.9, but using (a) 50 and (b) 100, 000 members from the distribution of Figure 2.3(c) to parameterise f. The grey contour shows the contour of Figure 2.9. The purple contour (a) is the result from using a KDE estimated using only 50 members for f.

The use of parameterisation to estimate the sampling uncertainty of quantiles is now investigated. In Figure 2.9, it was found that when estimating f with a KDE using the full 100, 000-member ensemble, Equation (3) gave a good approximation to the bootstrapped measurements of convergence, indicating that the convergence of uncertainty was well described by asymptotic theory. This was generally also the case for the height and rain model variables. The black contours of Figure 2.16(a) show the number of ensemble members required for a certain standard error for a range of quantiles as before, but now calculated using a Gaussian parameterisation for f. It uses 50 members from Figure 2.3(c) 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 line). 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 (3). 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 2.16(b)), 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.

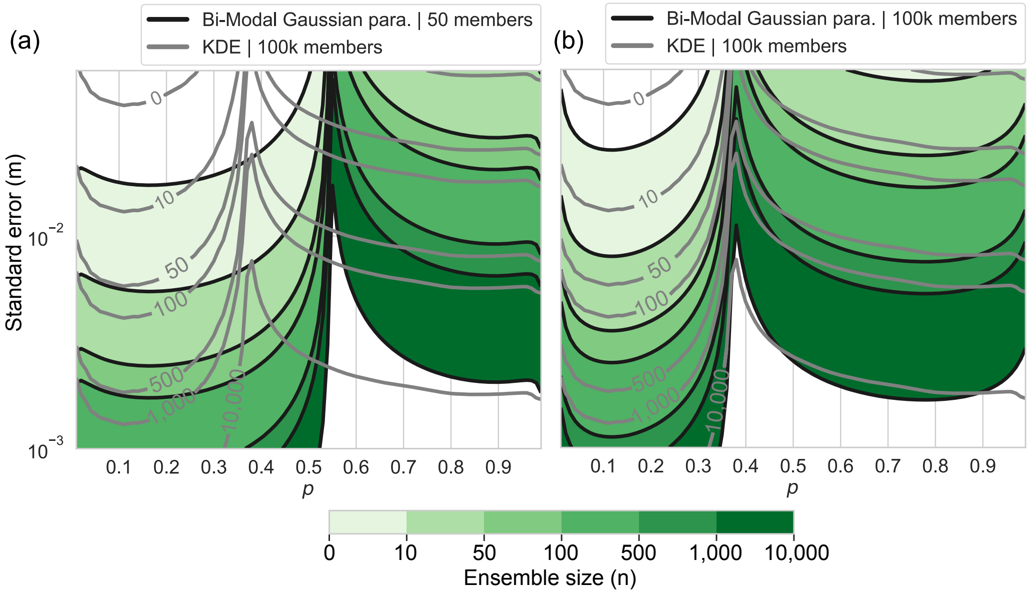

Figure 2.17: Black contours created as in Figure 2.10, but using (a) 50 and (b) 100, 000 members from the distribution of Figure 2.4(c) to parameterise f. The grey contour shows the contour of Figure 2.10.

A bi-modal Gaussian parameterisation of the height distribution shown in Figure 2.4(c) is used to estimate the convergence of sampling uncertainty in the quantiles (Figure 2.17).

Unlike non-parametric 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 $100,000$ are used to parameterise, the uncertainty estimate is closer to the KDE which used $100,000$ members for its estimation, but with slight over- and underestimation of uncertainty in regions. Only the bi-modal Gaussian parameterisation calculated using 50 members from the ensemble captures the decrease in uncertainty above the 0.96 quantile level.

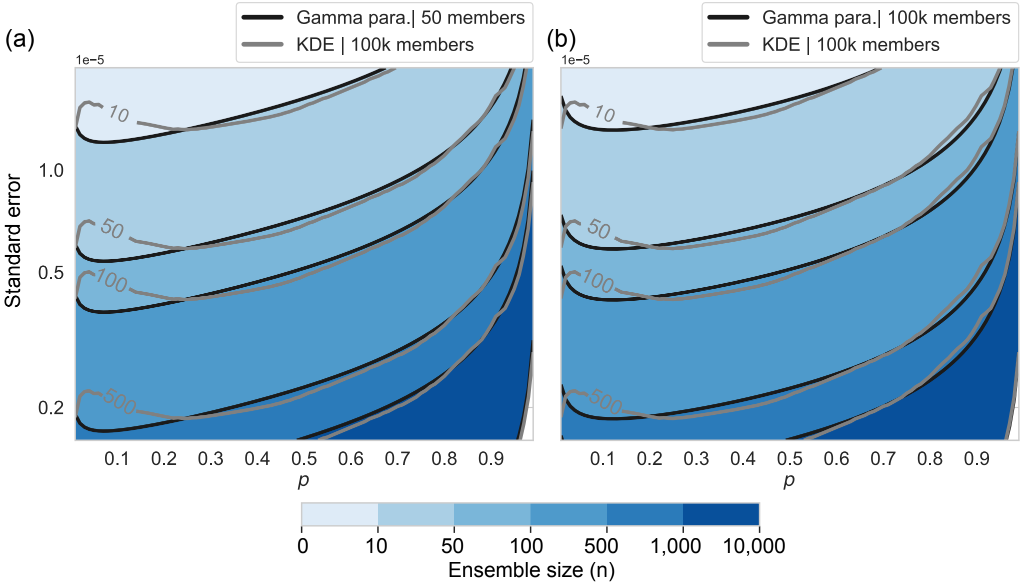

Figure 2.18: Black contours created as in Figure 2.11, but using (a) 50 and (b) $100,000$ members from the distribution of Figure 2.5(c) to parameterise f. The grey contour shows the contour of Figure 2.11.

Parameterising the rain distribution of Figure 2.5(c) with a Gamma PDF results in reasonable uncertainty estimates of the convergence of the sampling uncertainty of quantiles (Figure 2.18). 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 which 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 2.14, 2.15, 2.16, 2.17 and 2.18, 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 enough to estimate the convergence measure of the mean and variance, as well as quantiles near the median of uni-modal distributions. More members would be required for distributions with a multi-modal 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.

2.4 Summary and 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. Here an idealised prediction system which replicates the key processes of convection has been used to identify how sampling uncertainty of statistical properties converges with ensemble size in order to assess the relevance of asymptotic theory to estimating how large an ensemble should be.

The one-dimensional idealised prediction system developed was found to suitably replicate convective-scale forecast ensembles by comparing the ensemble distribution shapes for the three prognostic variables to corresponding quantities from a $1,000$-member full NWP ensemble. Shapes of these distributions over the 24 hour evolution in the free-run were found to fit into three categories: quasi-Gaussian, multi-modal, and highly skewed, as in our previous study (Craig et al., 2021). Also expected from previous work, the distributions became less Gaussian distributed in time, as anticipated due to the non-linear convective processes (Poterjoy, 2022; Kondo and Miyoshi, 2019; Kawabata and Ueno, 2020; Zhang, 2005; Legrand et al., 2016).

By creating a convergence measure, the sampling uncertainty was found to scale universally as $n^{-\frac{1}{2}}$ in the limit of large n. This applied to statistical properties including the mean, variance, quantiles between 0.05 and 0.95 as well as 0.999, the skewness and kurtosis. At what point 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 the marginals. However, due to the larger uncertainty associated with quantities from multivariate distributions, it would be expected that the absolute level of sampling uncertainty would be larger than for the uni-variate counterparts. It would also be expected that more members would be required until asymptotic convergence could be reached.

For the quantiles, the dependence of sampling uncertainty on distribution shape could be described by Equation (3), 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 to studies such as Leith (1974) which provides a specific number of members required to achieve sufficient precision in a certain aspect of an ensemble. Rather, the asymptotic convergence provides a scaling rule which 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 if 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. Another method that could have been used to reach the asymptotic regime with a small ensemble would have been to increase the effective ensemble size by implementing the neighbourhood method which was described in the Introduction.

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 if it would be worthwhile investing in expanding their current 50-member NWP ensemble to 100 members. They are particularly interested in their ability to accurately estimate the spread of temperature over Munich. To answer this question, they would like to calculate the convergence measure of the variance statistic for the temperature variable. The first thing they need to do is to check whether the asymptotic theory can be applied, by calculating the uncertainty in the convergence measure. They do this by bootstrapping their 50-member ensemble 100 times to obtain 100 distributions of length 50. With each of these distributions they then calculate the convergence measure. They see no divergence in the measures, similar to the green lines in Figure 2.12(a). It is in the asymptotic regime. This enables them to then visualise how the convergence measure will decrease as extra ensemble members are added to their 50-member NWP ensemble and hence how the accuracy of their estimate of the range of temperature over Munich will increase as more members are added. A knowledge of how many ensemble members to aim for in the future to obtain a certain level of sampling uncertainty can hence be calculated using a framework based on asymptotic theory.

Chapter 3

Asymptotic Convergence in Weak and Strong Forcing Convective Weather Regimes

3.1 Background

As laid out in the introduction, a major question of interest is how the convergence of sampling uncertainty with ensemble size changes depending on whether one is in the convective weather regime of weak or strong forcing. If differences are found, this would lead to different sizes of ensembles needed, depending on the regime.

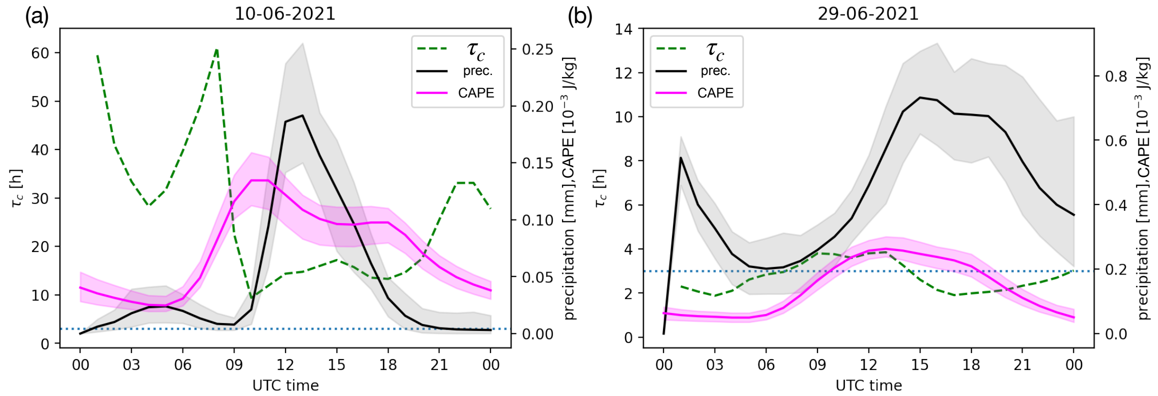

Figure 3.1: Daily evolution of the CAPE, convective timescale (τ_c) and precipitation on the (a) 10th and (b) 29th June 2021 during a period of weak and strong forcing respectively. Shading indicates the 95th CI. Simulated using data from 120-member ICON-D2 ensembles (Puh et al., 2023).

There are various measurements which can be made to differentiate between weak and strong forcing regimes. One such measurement is the convective adjustment timescale (τ_c) which was introduced in the Introduction. Along with this measure, the CAPE and precipitation can also be used to characterise regimes. An example of a weak and strong forcing regime is shown in Figure 3.1 using data from 120-member ensembles from ICON-D2, a full convective-scale NWP forecast (Puh et al., 2023)[3]. It shows the weak forcing day of the 10th June and the strong forcing day of the 29th June over Southern Germany. Beginning at 00:00 UTC in the weak forcing, it can be seen that the CAPE is relatively low, as well as the rain. Then as radiation 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 radiation 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 chapter.

As weak and strong forcing regimes have different behaviours as seen in Figure 3.1, it can be expected that the model variable distribution shapes from the two regimes also have differences. It has been seen in the previous chapter that the shape of the underlying distribution can influence the nature of how the sampling uncertainty converges with ensemble size (Tempest et al., 2023). As such, it is to be expected that differences in the distribution shape between the two forcing regimes could then propagate to differences in the convergence of sampling uncertainty as the ensemble becomes larger. Specifically, it might be expected 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 (3). 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, strong forcing would be expected to have a larger spread in its variables in the early hours and in the evening, which would mean a larger sampling uncertainty in the mean according to Equation (1). This would also apply to the standard deviation statistic described by Equa-

[3] I am a co-author, whereby I contributed data and analysis from the extended idealised ensemble developed in this Chapter

tion (2), however only in cases where the underlying distribution is normally distributed. It will be of interest to study these specific scenarios by analysing the distributions in weak and strong forcing situations as well as their complementary sampling uncertainty and how it converges with ensemble size.

A similar approach will be taken as in the previous chapter to investigate the difference in convergence of sampling uncertainty with ensemble size between weak and strong forcing. That is, an idealised model will be used to create large ensemble sizes. After it's success in our previous study (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 is 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 3.1. 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 (Figure 3.1), 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. The convergence measure from Chapter 2 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.

It is the aim of this chapter therefore to investigate how sampling uncertainty convergence with ensemble size differs between weak and strong forcing regimes and what implications this has for the ensemble size required. Section 3.2 will introduce the extended idealised ensemble which replicates weak and strong forcing. The distributions will then be analysed in Section 3.3.1 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 3.3.2 to ascertain whether the different forcing regimes do lead to different convergence behaviours. Summary and conclusions will then follow in Section 3.4.

3.2 Model and Methods

3.2.1 Extended idealised model

The idealised model from Chapter 2 (Tempest et al., 2023) is extended to allow for weak and strong forcing regimes. Figure 3.1 which uses data from a full NWP model will be used as a reference in this adaptation. Unless stated, the idealised model remains as previously described. 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:

$$\mathrm{CAPE}_e = (g\bar{h} - \bar{\phi_c})_e, \tag{10}$$

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 radiation 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:

$$\frac{d\phi_c}{dt} = S_{\mathrm{rad}} - S_{\mathrm{for}} + S_{\mathrm{con}}, \tag{11}$$

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

Radiation 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 radiation before 06:00 UTC and none after 18:00 UTC. Therefore:

$$S_{\mathrm{rad}} = \left\{ \begin{array}{cc} a_r \cos\left(\frac{2\pi t}{t_{\mathrm{total}}}\right), & \frac{t_{\mathrm{total}}}{4} < t < \frac{3t_{\mathrm{total}}}{4} \\ 0, & \text{otherwise} \end{array} \right\}, \tag{12}$$

where a_r has the value $0.000025\mathrm{m}^2\mathrm{s}^{-3}$. 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 the 29th 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:

$$A_{\text{for}} = a_f\left(1 + \sin\left(\frac{2\pi(t - T_{\text{shift}})}{T_{\text{for}}}\right)\right) \tag{13}$$

$$S_{\text{for}} = \left\{ \begin{array}{cc} A_{\text{for}}, & A_{\text{for}} > a_f \\ a_f, & \text{otherwise} \end{array} \right\}. \tag{14}$$

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 to balance it with the buoyancy terms of S_{rad} and S_{for}.

$$S_{\text{con}} = \left\{ \begin{array}{cc} \gamma_2\beta\frac{d\bar{u}}{dx}, & Z > H_r \text{ and } \frac{du}{dx} < 0 \\ 0, & \text{otherwise} \end{array} \right\} \tag{15}$$

where γ_2 is set to $-2000\mathrm{m}^2\mathrm{s}^{-2}$ and β is 0.1 as before.

To incorporate the diurnal cycle, whereby solar radiation 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:

$$C_{\text{stoc}} = \left\{ \begin{array}{cc} -\cos\left(\frac{2\pi t}{t_{\text{total}}}\right), & \frac{t_{\text{total}}}{4} < t < \frac{3t_{\text{total}}}{4} \\ 0, & \text{otherwise} \end{array} \right\}. \tag{16}$$

The addition of Equation (11) 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 radiation, 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 (5)) 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:

$$\tau_c = \frac{\text{CAPE}_e}{3600 \cdot S_{\text{con}_e}}, \tag{17}$$

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

3.2.2 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 in Chapter 2 was implemented, 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 EnKF 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 had the adaptations to the model as mentioned above. The free-runs for the weak and strong forcing ensembles were then ran for 24 hours, being initialised using the analysis' from the DA. The ensembles then had 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 (11) 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.

3.2.3 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 3.1).

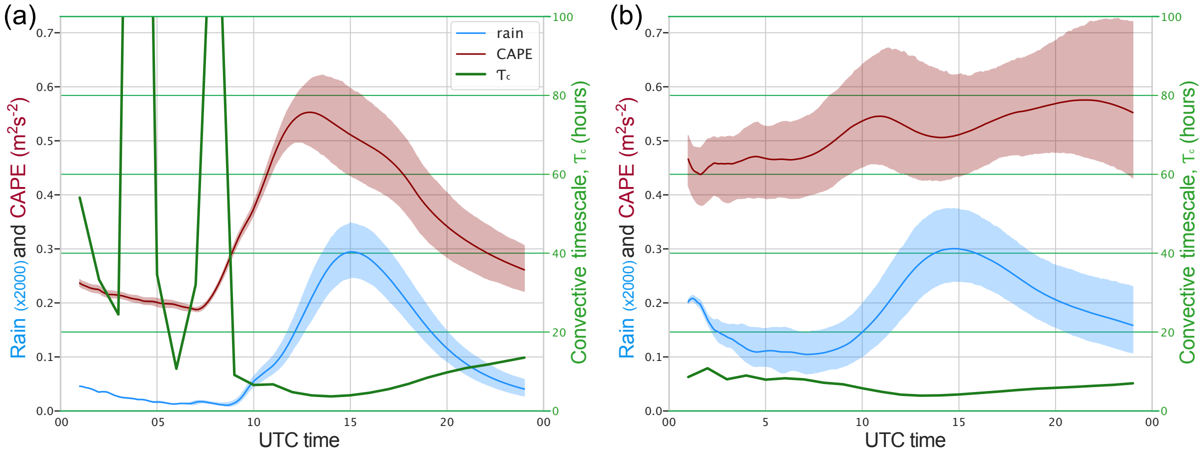

Figure 3.2: Time evolution of total rain in domain, CAPE and convective timescale in the (a) weak and (b) strong forcing free runs from the extended idealised ensemble. Shading indicates the 95% CI spread from the 5, 000-member ensemble. A simple moving average is used for the convective timescale.

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 3.2. As can be seen in the weak forcing case, the CAPE is low after 00:00 UTC and begins to increase after 06:00 UTC when the radiation 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 radiation 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 3.2 for the weak and strong forcing runs are in line with those computed from the ICON-D2 runs for the 10th and 29th June 2021 respectively (Figure 3.1). 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 3.2. 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 3.1 and 3.2 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 the previous chapter (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.

3.2.4 Analysis of neighbourhood distributions

To increase the effective ensemble size, the neighbourhood method is employed as it has been shown to be successful in previous convective-scale ensemble studies (Craig et al., 2022; Puh et al., 2023). 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.

3.3 Results and Discussion

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 weak and strong forcing regimes. The differences in distribution shape are furthermore expected to be important in explaining differences in the convergence measure (see Section 3.1). Expectations from the analysis of the distributions will therefore later be compared with the convergence measure calculated from the distributions.

3.3.1 Distributions from weak and strong forcing runs

Several characteristics of the distribution shapes are expected. First, it is expected that the three categories of distribution shape which are common to weather forecasting (quasi-Gaussian, multi-modal and highly skewed) (Tempest et al., 2023), 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 gathered 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.

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 3.1: Standard deviation of distributions from Figure 3.3.

To show the time evolution of the distribution shapes, Figure 3.3 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

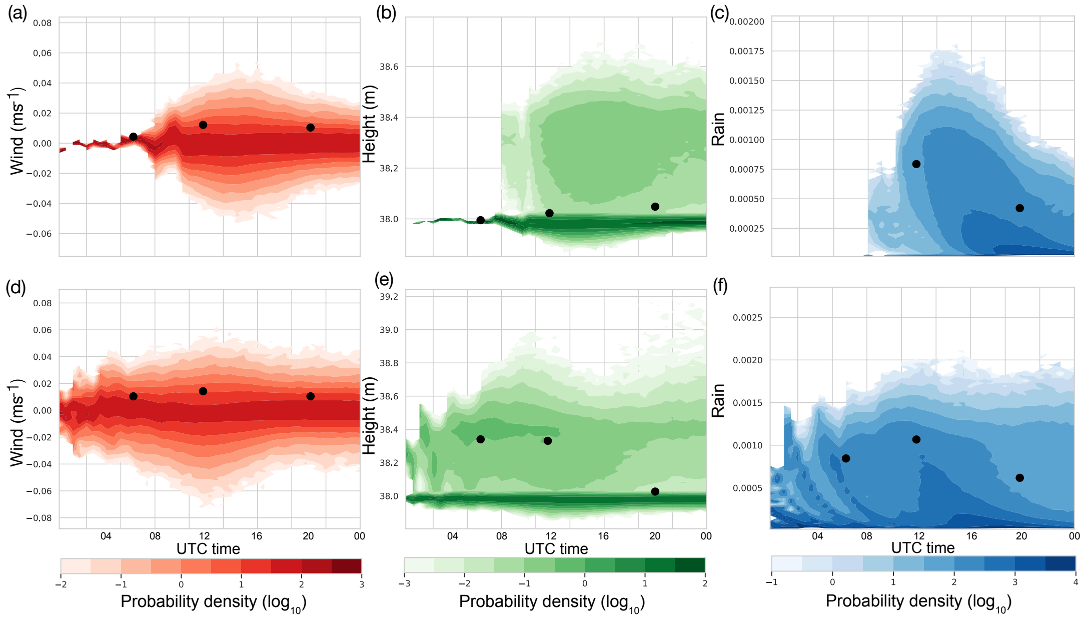

Figure 3.3: Contour histogram plots for the (a,d) wind, (b,e) height and (c,f) rain from the extended idealised ensemble over 24 hours in (a,b,c) weakly and (d,e,f) strongly forced regimes. Black dots indicate location of the 0.95 quantile. Neighbourhood of length 20km shown.

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 3.3, 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. Calculating the spread in Table 3.1, it is clear 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.

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 3.2: Probability density at the 0.95 quantile of the distributions in Figure 3.3.

The tails of the distributions from the weak and strong forcing runs are now compared in Figure 3.3. The probability density of the 0.95 quantiles are listed in Table 3.2 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 (Puh et al., 2023). 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 3.4 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 the 29th 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,

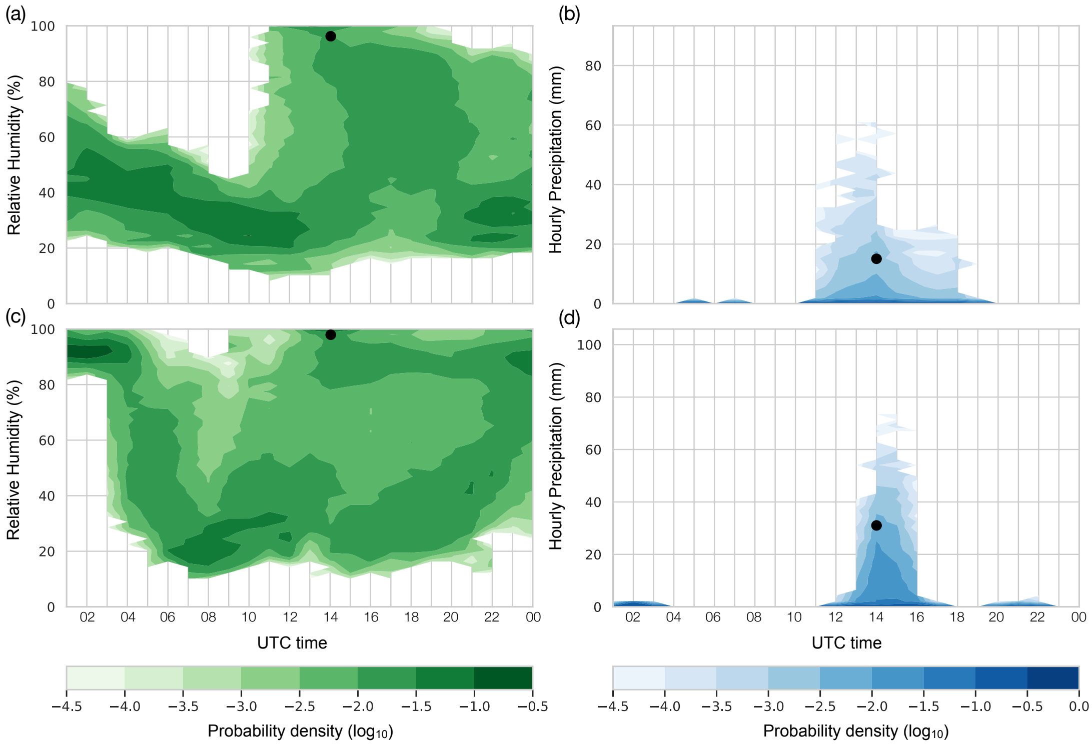

Figure 3.4: Contour histogram plots for the (a,c) relative humidity and the (b,d) hourly precipitation from the 120-member ICON-D2 ensemble over 24 hours on the (a,b) weakly and (c,d) strongly forced regime days of the 10th and 29th June 2021 respectively. Data used a circular neighbourhood of radius 10km. Black dots show location of 0.95 quantile at 14:00 UTC.

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 3.3 and 3.4, 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 3.4), it has been confirmed that even during periods of strong precipitation in the strong forcing, the weak forcing remains to the 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 variability. It is therefore expected that, according to Equation (1), 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. 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 (2), 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 (3) 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.

3.3.2 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 (1).

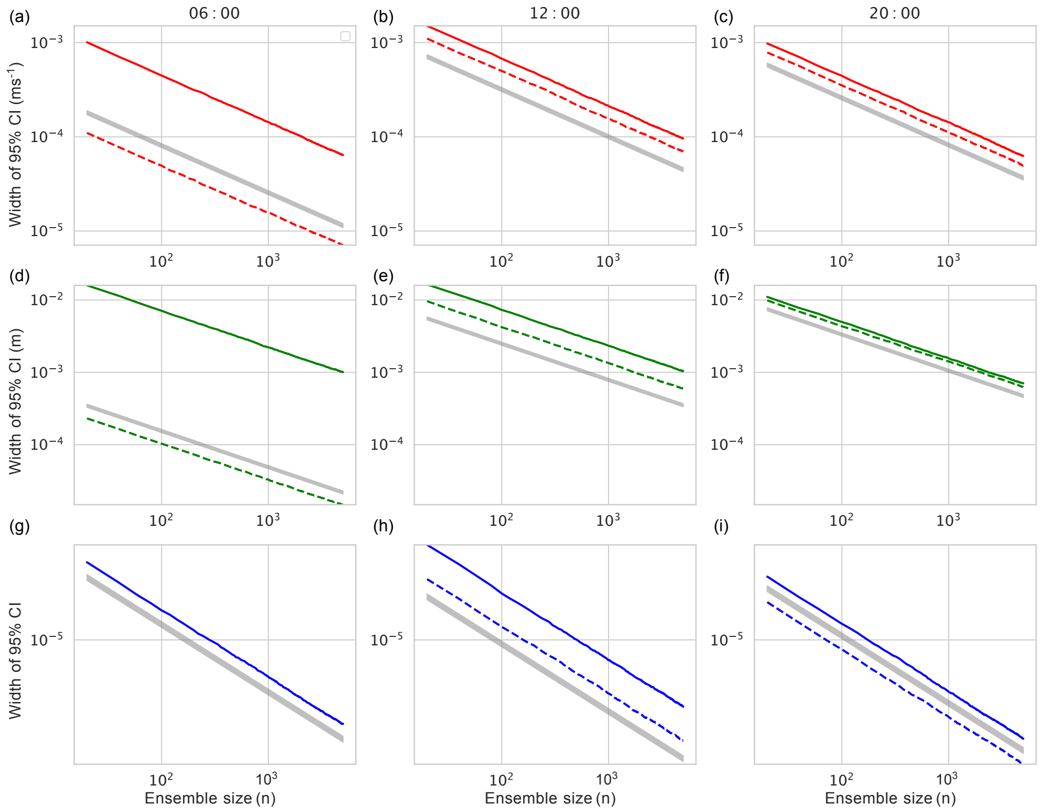

Figure 3.5: Convergence measure for the mean using data from the extended idealised ensemble with a neighbourhood of length 20km. Shows (a,b,c) wind (d,e,f) height and (g,h,i) rain. Each column is a different time point. Weak forcing is in the dashed coloured lines and strong forcing is in the solid 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.

Figure 3.5 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 contains no rain in weak forcing). As convection begins, these lines get closer together and this continues into the evening. This is expected according to Equation (1), 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 3.1) become closer together throughout the day, so does the sampling uncertainty of the mean.

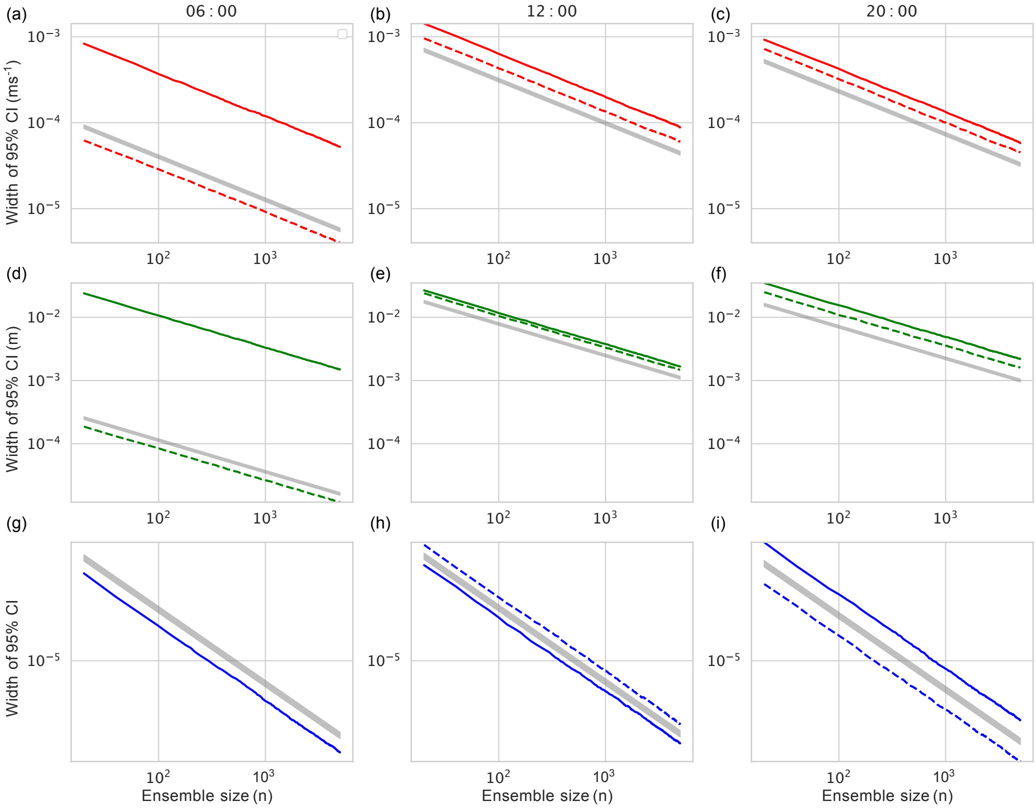

Figure 3.6: As in Figure 3.5 but for the standard deviation.

In a similar manner as the mean, the standard deviation convergence measure for the three model variables (Figure 3.6) is shown. 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 (2) 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 (2).

The convergence measures for the 0.95 quantile are shown in Figure 3.7 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 3.2, the strong forcing has the largest sampling uncertainty at all time points. Due to Equation (3), the

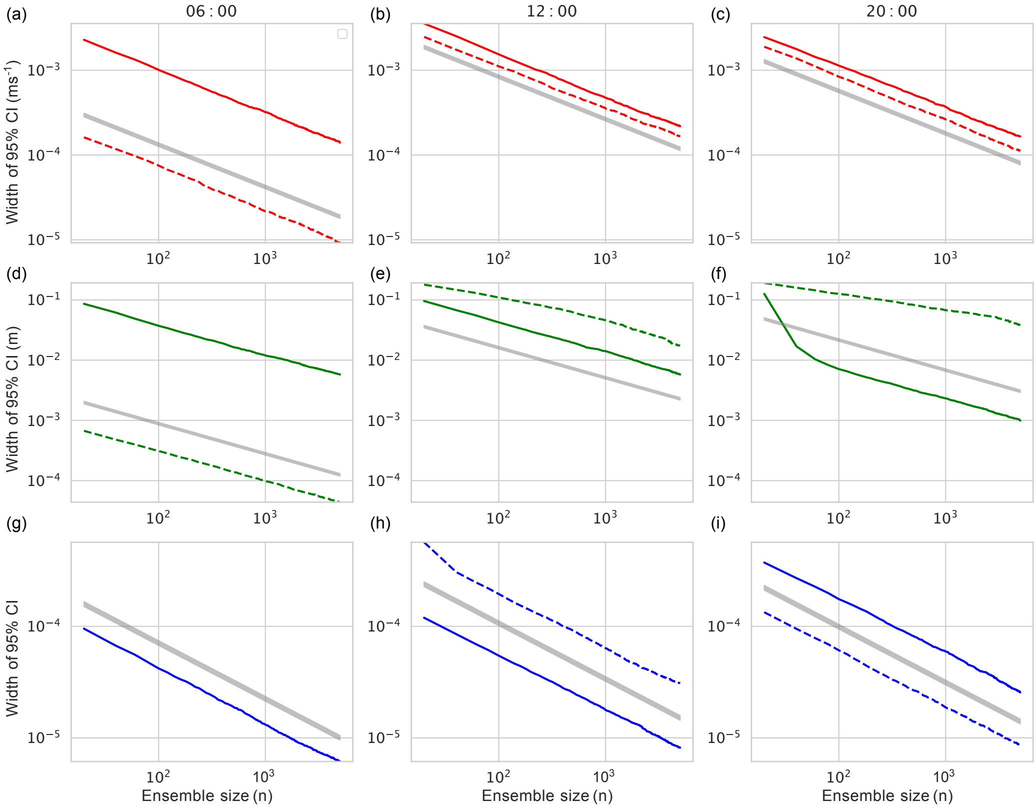

Figure 3.7: As in Figure 3.5 but for the 0.95 quantile.

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, the probability density at the 0.95 quantile is smallest for the weak forcing (Table 3.2) 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.

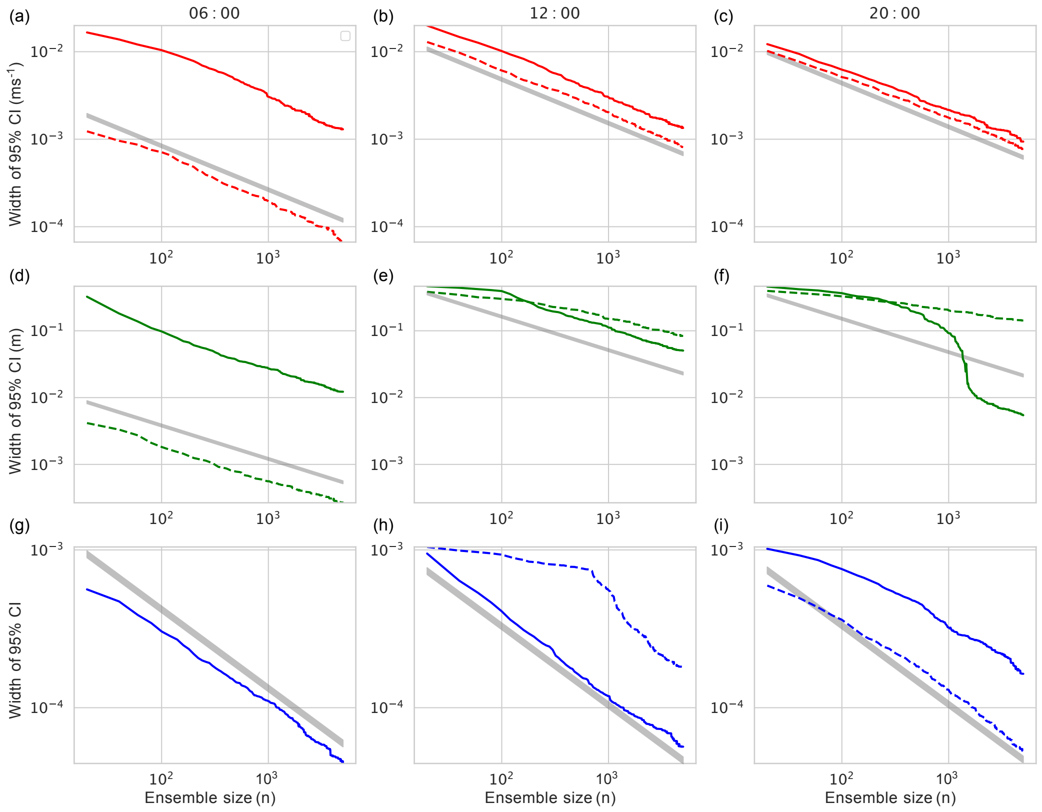

Figure 3.8: Convergence measure for the 0.95 quantile using data from the extended idealised ensemble from a single grid point. Shows (a,b,c) wind, (d,e,f) height and (g,h,i) rain. Each column is a different time point. Weak forcing is in the dashed coloured lines and strong forcing is in the solid coloured 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.

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 3.8 shows the convergence measure for the 0.95 quantile as in Figure 3.7 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 standard error 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.

3.4 Summary and Conclusions

Different weather regimes exist, each containing different types of weather. The question explored in this chapter was whether the ensemble size required to 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 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 Chapter 2 by adding in an extra time dependent equation for the constant geopotential term which was made to depend on the radiation, synoptic forcing and convection. The radiation 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 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 120-member full convective-scale NWP ensemble, ICON-D2, it was ascertained that distinct forcing regimes had been created by the extended idealised ensemble. Following Chapter 2, 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, Gaussian, multi-model and highly skewed were seen, in line with previous convective-scale studies. 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 (1), 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 (2) 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 (3). 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 differences highlighted of how sampling uncertainty converges with ensemble size between weak and strong forcing regimes indicates 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 the previous chapter, one can find the desired ensemble size for each regime by calculating the convergence measure. Say our forecaster with a 50-member ensemble is again interested in the spread of temperature over Munich and they want to know how many members they would require to reach a certain level of sampling uncertainty for days when there is weak forcing and for days when there is strong forcing. As before, they would calculate the convergence measure for the standard deviation (or variance) of the temperature and as seen in the previous chapter, it would likely be converging proportional to $n^{-\frac{1}{2}}$ with a 50-member ensemble. They would then extrapolate this 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 temperature distribution would have a larger spread and be Gaussian in shape, leading to it having a larger sampling uncertainty for any given ensemble size and as such require a larger 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, as before in Chapter 2. From this chapter 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.

Chapter 4

Asymptotic Convergence at the Synoptic Scale

4.1 Background

The topic of using different time and space scales in the prediction of the atmosphere, in particular the convective scale and the synoptic scale, was introduced in the Introduction. Ensemble size is a pressing question for both of these scales for ensemble forecasting and so it is valuable to ascertain whether the results of asymptotic convergence found previously for the convective scale are valid for the synoptic scale. If this is the case, a similar framework based on how the sampling uncertainty converges with ensemble size could be used to estimate the size of ensemble required for synoptic-scale probabilistic forecasts.

It has been seen in the previous two chapters that convergence of sampling uncertainty with ensemble size depends greatly on the shape of the underlying forecast variable distribution. From previous studies, it is generally expected that the shapes of distributions of forecast variables from synoptic-scale forecasts will fit well into the same three categories of Gaussian, multi-modal or highly skewed, that are seen in convective-scale forecasts. For example, temperature distributions can be modelled well by a normal distribution and generalisations thereof, including skewed normal and mixtures of normal distributions (Lakatos et al., 2022). Precipitation on the other hand can be modelled by a censored, shifted gamma distribution (Scheuerer and Hamill, 2015). For some forecast variables it is unsure how best it can be modelled, although it is thought in most cases that it has a distribution which belongs to one of the three categories of either Gaussian, multi-modal or highly skewed. For example the relative humidity, where it has been approximated as various symmetrical and skewed distributions (Tompkins, 2005). As many of the forecast variables from synoptic-scale models have similar distribution shapes as those from convective-scale models, it is expected that the characteristics of the convergence of the sampling uncertainty with ensemble size will also be similar.

For the investigation of whether asymptotic convergence in the sampling uncertainty occurs for synoptic-scale data, a model with the suitable scales must be chosen. Furthermore the ensemble should be tuned to have the appropriate amount of spread. The Integrated Forecast System (IFS) from ECMWF is chosen, which has a horizontal resolution close to 18km and a lead time of up to 15 days. The IFS ensemble is moreover well developed and used operationally. It is composed of 50 members with perturbed initial conditions, 1 control member and 1 high resolution run. A multitude of model variables are archived and available for analysis. The temperature at 2m and 500hPa, the relative humidity and the accumulated precipitation will be analysed in this chapter for the reason that they give a sample of each of the three categories of distribution shape and will allow for comparison with our convective-scale results Puh et al. (2023).

It has been hypothesised that it is expected that the sampling uncertainty convergence properties will be similar for synoptic-scale data and convective-scale data due to the distribution shapes. For a 50-member synoptic-scale ensemble it is therefore conjectured that the sampling uncertainty of the mean, most standard deviations and non-extreme quantiles will be converging proportional to $n^{-\frac{1}{2}}$, where n is ensemble size. Whereas extreme quantities which require more resolution at the tails of the distribution such as the 0.95 quantile, will not converge and need more ensemble members. A statistic which is commonly calculated operationally with synoptic-scale forecasts is the Extreme Forecast Index (EFI), which is used to identify potentially extreme events (ECMWF, 2023). It is uncertain whether a 50-member ensemble is large enough to see asymptotic convergence in other forecast statistics such as this one, if it does indeed converge proportional to $n^{-\frac{1}{2}}$ in the limit of large n. Application of the neighbourhood method to increase the effective ensemble size is not expected to be as effective with the larger-scale data as for the smaller-scale data. This is due to the higher correlated nature of the grid points, leading to them being less independent. It is of interest whether this method can still be effectively used however, in some capacity to increase the effective ensemble size.

In this chapter the IFS from ECMWF will be analysed to investigate whether the asymptotic sampling uncertainty convergence seen in convective-scale forecast data also occurs in synoptic-scale forecasts. The IFS and the relevant methods are described in Section 4.2. In the results section of 4.3, the distributions from the model are first analysed for four model variables to check how similar the distribution shapes are in the synoptic scale compared to the convective scale. Then the convergence measure, introduced in Chapter 2 which measures the convergence of sampling uncertainty with ensemble size, will be calculated for various statistics, including the EFI. This convergence measure will be analysed to ascertain how many members are needed for scaling proportional to $n^{-\frac{1}{2}}$, as well as the effect of the neighbourhood size and how similar these results are from the convective scale. Finally in Section 4.4 it will be concluded whether a similar framework to calculate ensemble size can be used in the synoptic scale as for the convective scale.

4.2 Model and Methods

4.2.1 ECMWF model

The operational ensemble forecast of the ECMWF IFS consists of the ECMWF Ensemble of Data Assimilations (EDA) (Isaksen et al., 2010; Lang et al., 2019; ECMWF, 2021a) and the Ensemble Forecast (Palmer et al., 1992; Molteni et al., 1996; ECMWF, 2021c). The IFS is a spectral, hydrostatic model and uses terrain following pressure coordinates (Rodwell and Wernli, 2022). Both the EDA and Ensemble Forecast consists of 50 ensemble members and 1 control member with a 12 minute time step. The horizontal grid has a resolution close to 18km and there are 137 levels vertically.

The EDA system uses a multi-resolution incremental 4D-Variational method, a type of DA which involves minimising a cost function using information from observation operators and adjoint and tangent-linear versions of the non-linear forecast model (ECMWF, 2021a). High and low resolutions are used to make the process more efficient. The observations which are screened and corrected using the Variational Bias Correction (Dee, 2004) are organised in time-slots every 30 minutes and are randomly perturbed to simulate observation uncertainty. The assimilation windows are 12 hours in length, in which time the distance between the model-trajectory and information from the background and observations is measured by the cost-function to then calculate an analysis as well as its associated uncertainty. The uncertainty of the analysis, which is the initial condition uncertainty, will then be a combination of the observation and background uncertainty. The ensemble which then is used to initiate the forecast will be constructed from this initial condition uncertainty.

The IFS employs various parameterisation schemes which simplify complicated processes in the atmosphere. Examples include a large-scale cloud and precipitation scheme (Tompkins and Janisková, 2004) as well as a control for gravity waves (ECMWF, 2021b). For representation of errors and (sub-grid-scale) uncertainties, the Stochastic Perturbation to Physical Tendencies (SPPT) parameterisation (ECMWF, 2021c) is used. To further increase spread in the Ensemble Forecast during the first 2 days, singular vectors (Molteni and Palmer, 1993; Leutbecher and Lang, 2014) are added, which perturb the initial conditions provided by the EDA.

4.2.2 Ensemble forecast

The operational forecast which is analysed in this chapter begins at 00:00 UTC on the 10th June 2021. The forecast is run for 15 days and the model variables of the temperature at 500hPa, the temperature at 2m, the relative humidity and the total precipitation are extracted.

4.2.3 Data analysis

Neighbourhood distributions

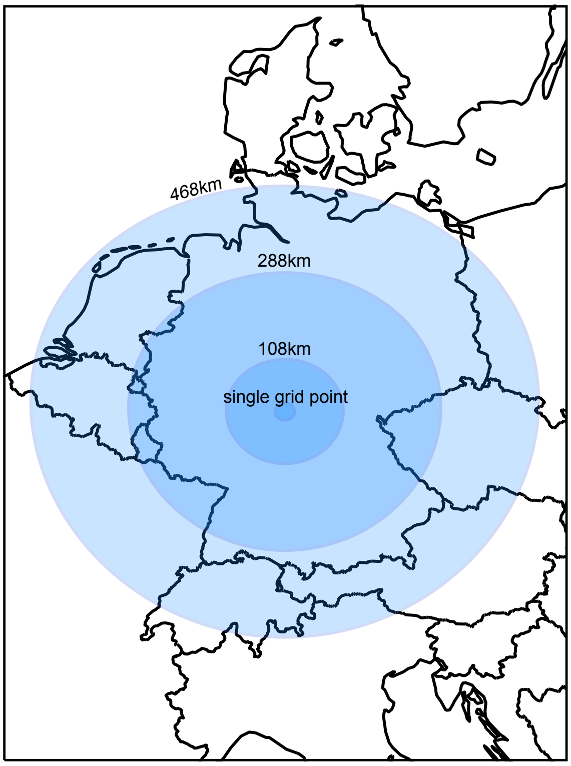

Figure 4.1: Black lines show geographical outline of central Europe. Filled blue circles show regions used to create neighbourhoods of various sizes. Centre grid point used for each neighbourhood is at (50.39958N, 9.686247E).

Similar neighbourhoods are created as in our studies of Craig et al. (2022) and Puh et al. (2023), the centre point of which were taken to be at the coordinates of (50.39958N, 9.686247E). Three neighbourhoods were then created, with radii of 108km, 288km and 468km, as seen in Figure 4.1.

Model Climate

The model climate, otherwise known as M-climate, is the climatological distribution of a forecast variable and is used in the calculation of select statistics. It is the distribution a forecast variable achieves when it contains no more predictability. As the climate is seasonal and varies in location, it is calculated for a specific time of the year and specific location.

For the preparation of the model climate, three sets of nine consecutive re-forecast sets are used. Of the nine sets, the 5th is set to be corresponding to the preceding Monday or Thursday (days on which the reforecast is initiated) which is closest to the relevant date of interest. Each of these sets consists of an 11-member ensemble (1 control and 10 perturbed members) with data for the previous 20 years at the same date. The distributions are then extracted for the relevant lead time. Although in practice only one starting year is chosen, in this thesis the process is carried out three times: with starting years of 2019, 2020 as well as 2021. The combination of these ensembles then comprises the model climate distribution of size (3 starting years·9 sets·11 members·20 years) = 5940 members.

EFI statistic

The EFI is calculated by comparing the ensemble distribution of a forecast variable to the model's climatological distribution for the chosen location, time of year and forecast lead time. It can have values between −1 and 1. The more this value deviates from 0, the more unusual the forecast distribution is. The EFI is calculated by the following:

$$\mathrm{EFI} = \frac{2}{\pi} \int_0^1 \frac{Q - Qf(Q)}{\sqrt{Q(1-Q)}}\, dQ, \tag{18}$$

where $Qf(Q)$ denotes the proportion of ensemble members from the forecast situated at the Q quantile of the model climate (ECMWF, 2023). The denominator provides more weight to the extremes of the model climate. Note that if the ensemble distribution is outwith the values of the climatological distribution, these are not taken into account in the calculation. As the model climate dataset is much larger than the ensemble distribution, the estimated model climate distribution is assumed to be the true distribution and the sampling uncertainty discussed in this chapter will relate purely to the ensemble forecast distribution.

Note that in our calculations of the EFI, although they were checked by multiple scientists, the EFI computed matches the ECMWF value only within a factor of 10^{-2}. This could be due to a number of differences including the specific files used to create the model climate as well as the bin size used to create the CDF in the calculation of the EFI.

4.3 Results and Discussion

4.3.1 Distributions from the synoptic scale model

As a first step in analysing the sampling uncertainty convergence behaviour of the synoptic-scale data from ECMWF, the 50-member forecast distributions are examined, since it has been seen in previous chapters that the distribution shape has an important role in determining the nature of the convergence of sampling uncertainty. It is of interest to see whether the distributions from the ECMWF model match the shape of distributions from previous synoptic-scale and convective-scale studies. As the distribution shape varies throughout the forecast, various time points will be looked at and it will be expected that the evolution will follow the conceptual framework proposed by our previous study (Craig et al., 2022). In addition, the effect of neighbourhood size on the distribution shape will be analysed. It is expected that above a radius of approximately 100km the distribution shape will change, because different orographies and synoptic weather conditions would likely be included.

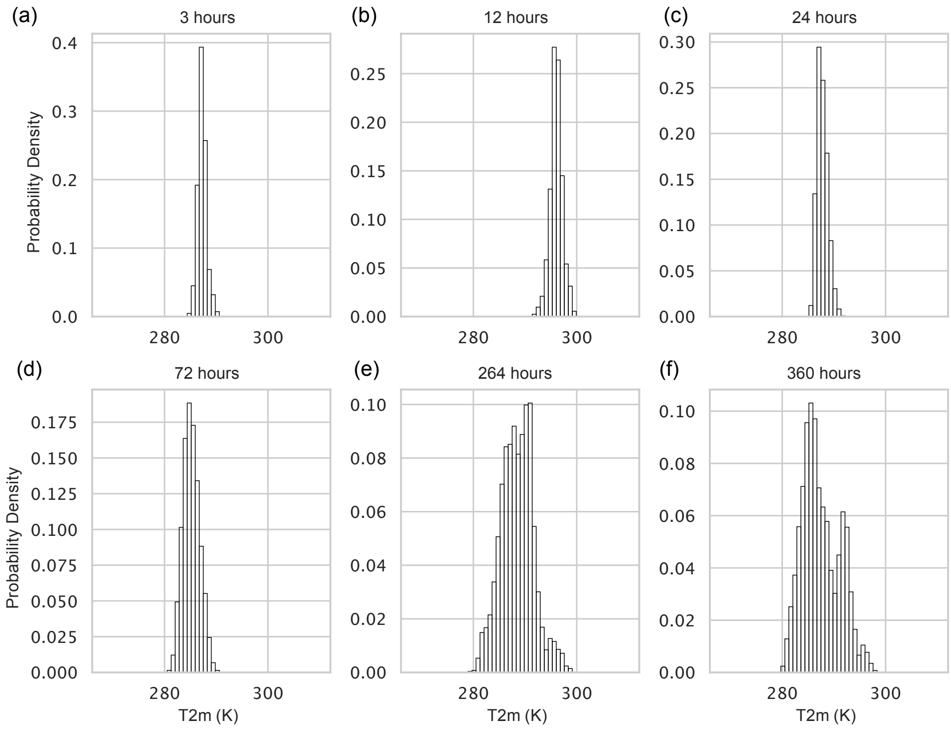

Figure 4.2: (a-f): Evolution of distributions in time of the temperature at 2m using the 50-member IFS ensemble with a 108km radius neighbourhood. Title of each subplot shows lead time of forecast.

Figure 4.2 shows distributions of the temperature at 2m for various lead times with a neighbourhood of radius 108km. This neighbourhood is used so that there are as many members in the histogram as possible before the distribution shape was seen to change significantly. The distribution at each time point is shaped by either one, or a combination of, Gaussian distributions, as expected (Lakatos et al., 2022). The expected distribution shapes according to synoptic-scale studies are also seen for the other three model variables inspected, see Appendix Figures B. 6.2, 6.1 and 6.3. The relative humidity and temperature at 500hPa can be similarly modelled by a selection of Gaussian distributions and the total precipitation can fit a gamma distribution well. It is therefore seen that the shapes from the ECMWF model are as expected following synoptic-scale studies and also fit into the same three categories as seen in convective-scale ensembles.

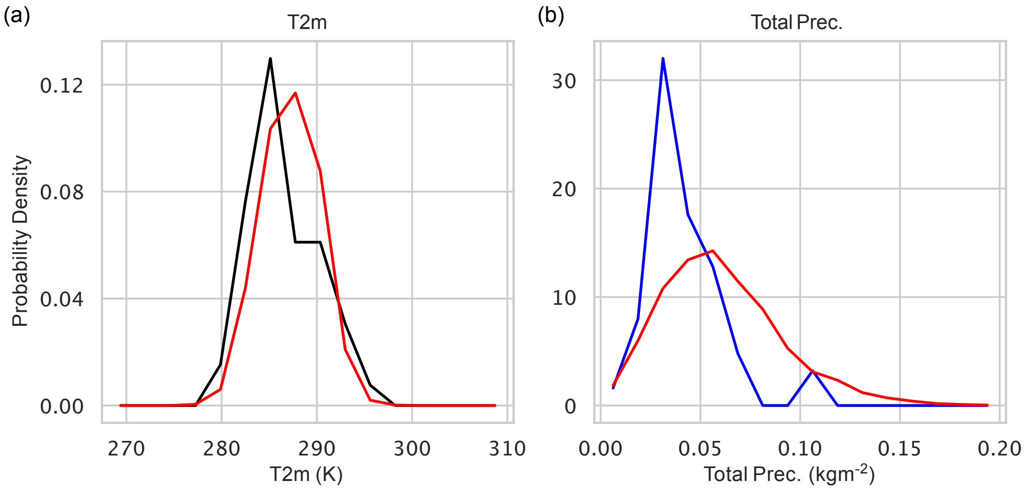

Figure 4.3: Single grid point distribution at 360 hours of free forecast for the (a) temperature at 2m and the (b) total precipitation using the IFS ensemble. Lines plotted connect each of the 16 histogram bins. The model climate distribution from reforecasts is overlaid in red.

Over the 15 days of forecast, the forecast distributions evolve their shape. At the beginning of the forecast, at 3 hours, the distributions for all four model variables are a relatively narrow Gaussian. In the case of the temperature at 2m (Figure 4.2) this Gaussian shifts its mean to higher values and then shifts back to lower temperatures later on in the day. During the first 12 hours of forecast, the temperature at 2m is likely increased due to radiation, and then likely decreases in the afternoon due to precipitation evaporating and less radiation being absorbed from the sun. During the rest of the forecast the distribution broadens out, obtaining a multi-modal Gaussian at the end of day 15. It is questionable whether the distribution at 360 hours is the climatological distribution. To test this, the distribution from the temperature at 2m and total precipitation forecast was compared with their model climate distribution and can be visualised in Figure 4.3. To aid visualisation, lines are plotted which connect each of the 16 histogram bins. As can be seen by comparing the single grid point distributions from the forecast to the model climate created from reforecasts, the climatology has not been reached after 360 hours of free forecast. The evolution of distribution shape in time has therefore been observed to follow our conceptual framework (Craig et al., 2022) which was previously seen to apply for convective-scale forecasts. Whereby the distributions from the synoptic-scale ensemble begin with a constrained distribution from the DA, which then develops varied shapes (mainly in the form of shifts in the case of the 2m temperature). Like the convective-scale forecasts however, the final stage of reaching the climatological distribution is not reached after 15 days of forecast.

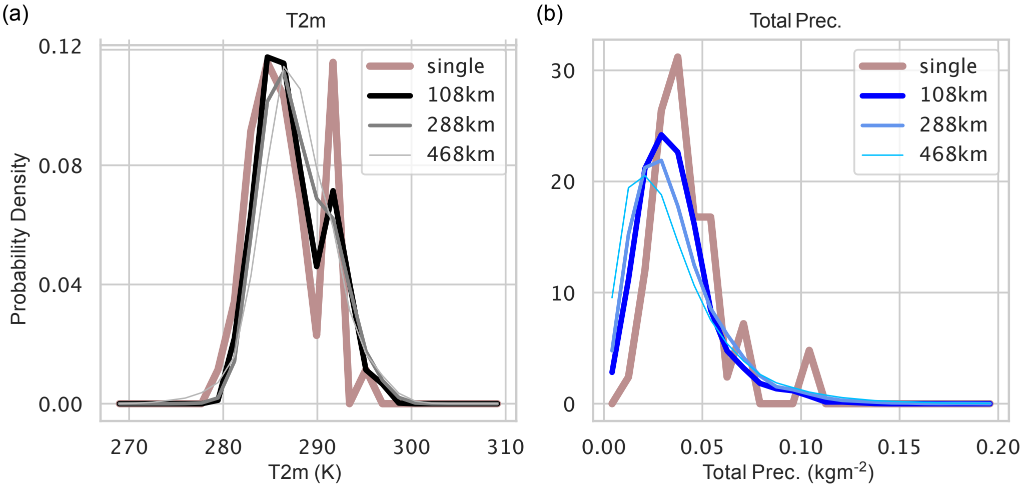

Figure 4.4: Effect of neighbourhood size on the distributions of (a) temperature at 2m and the (b) total precipitation at 360 hours of forecast lead time using the IFS ensemble. Lines plotted connect each of the 24 histogram bins. “single” in legend refers to the single grid point and values in kms refers to the radii of neighbourhood regions.

Figure 4.4 shows distributions of the temperature at 2m and total precipitation for 3 neighbourhood sizes and the single grid point at 360 hours of forecast lead time. Although the distributions are not matching exactly, as shown by the 2m temperature bimodality seen in the 108km radius neighbourhood not appearing in the larger neighbourhood sizes, they are more similar than at the beginning of the forecast and this was also the case for the two other variables (not shown). As expected, the larger the neighbourhood, the larger the spread. This is particularly obvious for the temperature at 2m where the largest neighbourhood has a longer tail on the left side of the distribution. This occurs because the largest neighbourhoods encompass more regions and so have a larger range of values. The other variables again had similar characteristics in their neighbourhood distributions (not shown). As the larger neighbourhood sizes do not change their distribution shape significantly, it is possible that in this forecast case and lead time, the neighbourhood method could be useful to expand the ensemble.

It has been shown that the distributions from the ECMWF data are as to be expected according to previous studies of synoptic-scale data. Furthermore these shapes of distribution fit into the same three categories as seen in convective-scale data, and their evolution in time follows a similar conceptual framework. Since the distribution shapes are similar, this provides a good basis to assume that our hypothesis will hold that there will be similar behaviour of the sampling uncertainty convergence with ensemble size in the synoptic-scale data as with the convective-scale data. Furthermore, at later lead times the distribution shapes from neighbourhoods with radii larger than 100km are similar to those with smaller radii. This suggests that the neighbourhood method could be more useful than initially expected at allowing the asymptotic convergence of sampling uncertainty to be seen in more extreme statistical quantities otherwise hidden with a small 50-member ensemble.

4.3.2 Sampling uncertainty convergence

It has been seen that the distribution shapes throughout the 15 days of the ECMWF forecast fit into the same three categories of distribution shape as seen in convective-scale models and as such it can be expected that similar convergence behaviour of the sampling uncertainty will apply to the synoptic scale as for the convective scale. To assess this, the convergence measure for a selection of standard statistics has been computed for the same four model variables as analysed for the distributions at 360 hours. This time point has been chosen due to it containing a variety of distribution shapes. A single grid point is further analysed as it will allow for it to be easily seen how many members are needed for asymptotic convergence proportional to $n^{-\frac{1}{2}}$ to occur, if it does at all. It is noted that the distribution shape of the single grid point and the neighbourhood with radius of 108km have similar shapes at 360 hours. Convergence proportional to $n^{-\frac{1}{2}}$ is expected with the 50-member ensemble for the statistics of the mean and in most cases for the standard deviation and the least extreme quantiles. The neighbourhood method may be of help to then artificially increase the ensemble size if convergence asymptotically has not been obtained although this may not function well due to the higher correlated nature of synoptic-scale data.

Figure 4.5 shows the convergence measure for four variables (columns) and various statistics (rows) using ensemble data from a single grid point at 360 hours into the forecast. The thicker grey line in the background is converging proportional to $n^{-\frac{1}{2}}$, in order to observe whether the sampling uncertainty of a forecast variable is converging asymptotically. Concentrating first on the mean (top row) and the temperature at 500Pa (first column), one sees that the convergence measure converges asymptotically with less than 10 members. This is also the case for the temperature at 2m, relative humidity and total precipitation. For the standard deviation convergence, the model variables apart from the total precipitation converge after less than 50 ensemble members. More members would be required in the total precipitation. This is likely due to the strongly skewed distribution shape of the total precipitation, as extra members could add to the tail, varying the standard deviation significantly. The convergence measure of the mean and standard deviation at other time points are similar. They have mostly converged apart from a case of a very highly skewed distribution with a large concentration at zero whereby it's standard deviation hasn't converged with 50 members (not shown). This follows our expectations

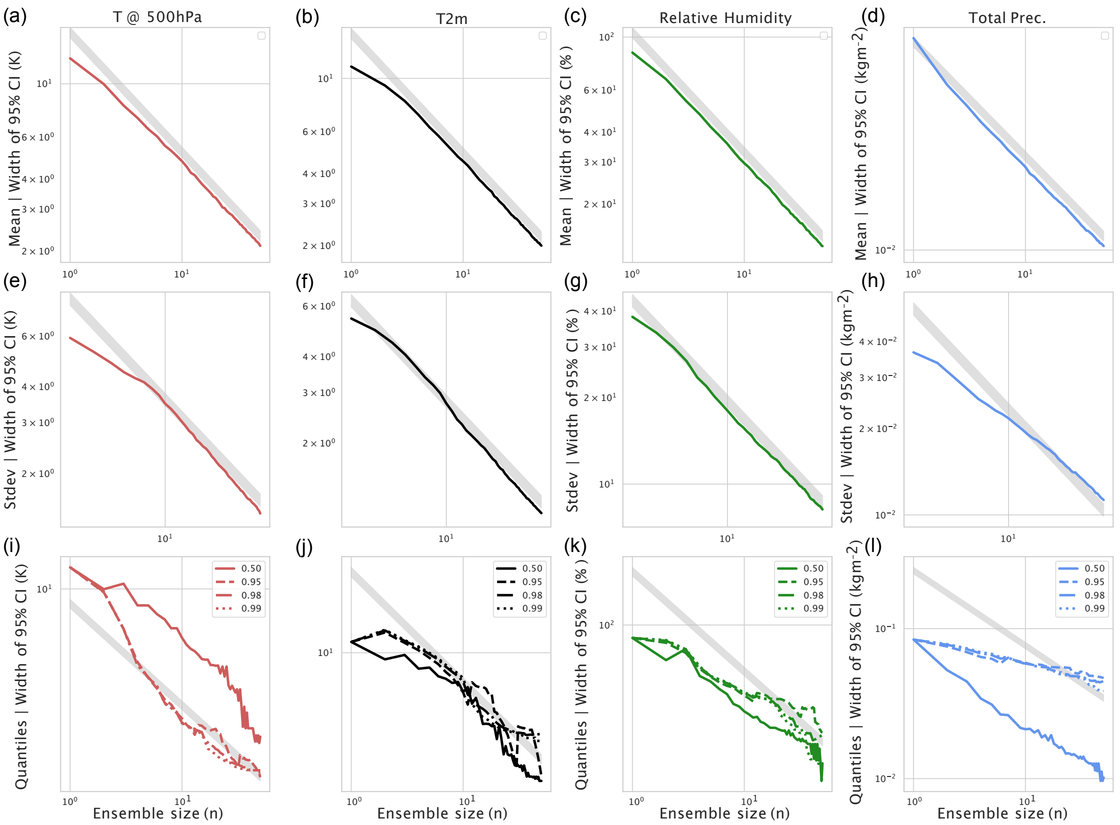

Figure 4.5: Convergence measure using IFS ensemble data from a single grid point and four model variables (columns) at 360 hours of forecast lead time. Rows contain different statistical quantities. From top to bottom they are the (a-d) mean, (e-h) standard deviation and (i-l) multiple quantiles. Grey line in background is converging proportional to $n^{-\frac{1}{2}}$ and shows the expected slope for asymptotic convergence.

from the convective scale whereby the mean often converges asymptotically with less than 50 members and the standard deviation needs between 50 and 100.

None of the quantiles on the bottom row of Figure 4.5 (the line style depicts the quantile level, going from $p = 0.5$ to 0.99) are converging proportional to $n^{-\frac{1}{2}}$, indicating that none have reached the asymptotic regime with only 50 ensemble members. This includes the 0.5 quantile, otherwise known as the median. The median has not converged but the mean likely has because if one adds another ensemble member to the 50-member ensemble, the median can vary substantially more than the mean since the mean would include the extra point into the overall average whereas there could be a large gap to the new median point in the data. A “zigzagging” is additionally seen in the convergence measure of the quantiles. This is due to the small ensemble size and is also seen in convective-scale data from ICON-D2 (Puh et al., 2023). The convergence measure for the quantiles from the three other model variables at other time points are similar: they have not asymptotically converged with a 50-member ensemble (not shown). This is broadly in line with studies from the convective scale whereby it was seen that for quantiles at unpredictable parts of distributions (e.g. tails and troughs), more members are needed for convergence proportional to $n^{-\frac{1}{2}}$. Asymptotic convergence could have been expected however for the more central quantiles as in Figure 2.8 of Chapter 2. Further investigation is required to understand this behaviour.

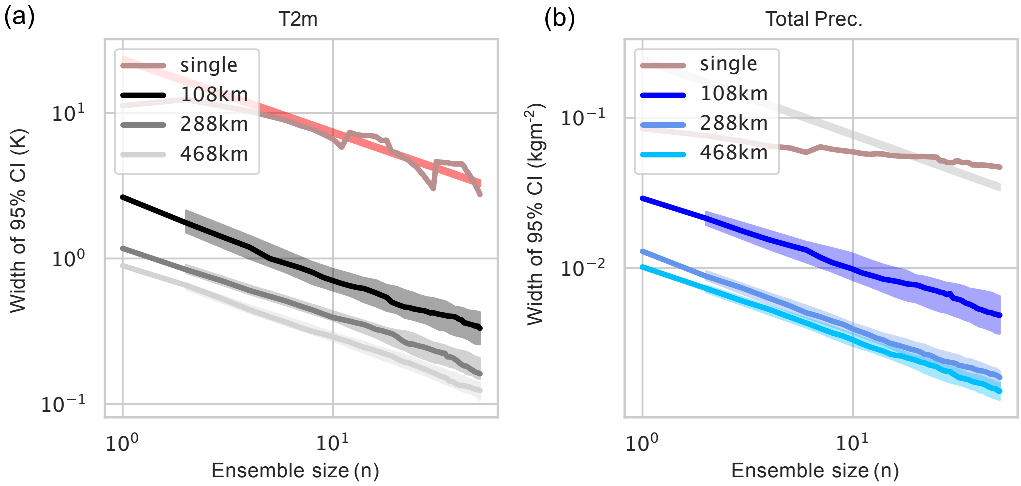

Figure 4.6: Convergence measure for the 0.95 quantile statistic for the (a) 2m temperature and (b) total precipitation at 360 hours of forecast lead time. Red and grey line in background of single grid point are converging proportional to $n^{-\frac{1}{2}}$. Convergence measure for various neighbourhood sizes are in addition plotted. "single" in legend refers to single grid point and values in kms refers to the radii of neighbourhood regions. Shading behind the neighbourhood convergence measures are the 95% CI of the uncertainty of the convergence measure.

The neighbourhood method is used in an attempt to enlarge the ensemble artificially in order to achieve asymptotic convergence in the sampling uncertainty in select quantiles that did not converge with a single grid point in Figure 4.5. As a quantity which did not converge with 50 members, the 95th percentile of the 2m temperature and total precipitation is chosen. The convergence measure is calculated for this statistic for neighbourhood radii of 108, 288 and 468km and is shown in Figure 4.6. The Figure shows that as the neighbourhood size increases, for both model variables, the convergence measure shifts to smaller values, decreasing the magnitude of the sampling uncertainty. It is perhaps surprising that the neighbourhood method appears to work due to the relatively correlated nature of the grid points over large regions. However, the neighbourhood regions chosen are quite large, and the shapes of the largest neighbourhood distributions are not too dissimilar to the smaller neighbourhood distributions (see Figure 4.4) indicating that the statistical properties are not too indifferent and so this likely leads to the neighbourhood method being effective in this case.

The uncertainty of the convergence measure is further assessed in order to determine which of the neighbourhood lines in Figure 4.6 have converged asymptotically. The single grid points are not tested as they are clearly not scaling proportional to $n^{-\frac{1}{2}}$. The same method as in Section 2.3.3 is carried out whereby the convergence measure is calculated many times using bootstraps of the original ensemble to calculate new convergence measures. One hundred of these calculations were carried out, from which a 95% CI was computed. This is shown by the shading behind the convergence measure lines. Clearly, for both model variables, the width of the CI decreases as the size of the neighbourhood region increases. Furthermore, at the smaller neighbourhood sizes one sees the CI slightly diverging at larger ensemble sizes, indicating that it has not converged asymptotically. For example the total precipitation's neighbourhood of radius 108km. With a neighbourhood region of radius 469km, the convergence measure has asymptotically converged for both model variables. From computing the uncertainty of using the neighbourhood method to increase the effective ensemble size, it has been seen that in this case where large neighbourhoods don't lead to significantly different distribution shapes, that it can be an accurate and cheap method to artificially increase the ensemble size to achieve asymptotic convergence.

From investigating how sampling uncertainty decreases with ensemble size using data from the ECMWF forecast, asymptotic convergence was seen in synoptic-scale data in the statistics of the mean, standard deviation and quantiles when the neighbourhood method was employed. Increasing the effective ensemble size using the neighbourhood method was seen to be more useful than expected for being able to reach asymptotic convergence with a limited ensemble size. This could be because of the statistically similar distributions up until large neighbourhood sizes in the case selected. It has been seen as a whole that the previous conclusions from the convective scale (Chapters 2 and 3) also apply to the larger time and space scale data from the ECMWF forecast. A further question which is of interest in particular for this ECMWF dataset is whether asymptotic convergence can be seen in statistics commonly used in weather forecasting. As such, the sampling uncertainty convergence of the EFI will now be investigated.

EFI

In operational settings, which is where data from the IFS ensemble is often employed, statistical quantities other than the mean, standard deviation and quantiles are often also used. One of these is the EFI. As the EFI, which is calculated by Equation (18), can be thought of as the sum of iid random variables multiplied by a factor, it is expected to follow the CLT. This would mean that the sampling distribution of the EFI would be normally distributed and as such will converge asymptotically with enough ensemble members. It will now be explored whether asymptotic convergence of the EFI does occur with the 50-member IFS ensemble.

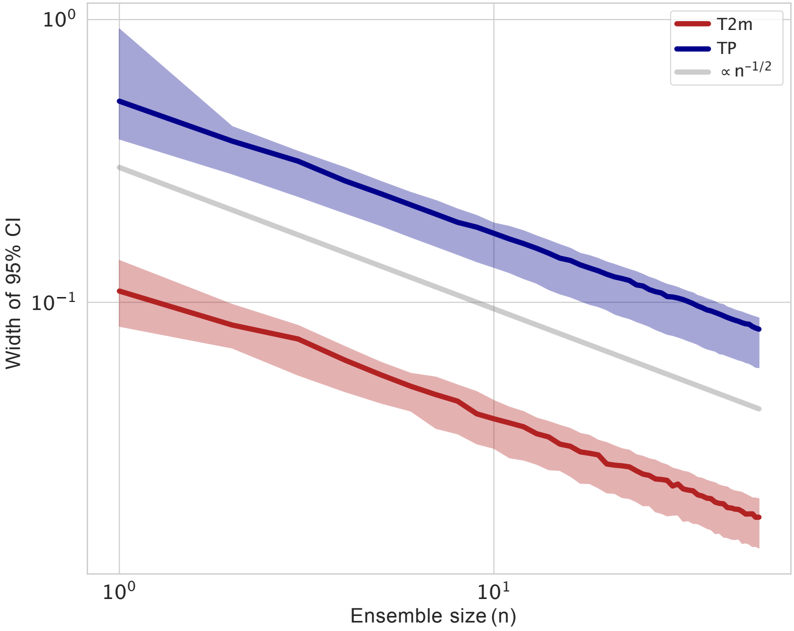

Figure 4.7: Convergence measure of the EFI for the (a) 2m temperature and the (b) total precipitation shown by red and blue lines respectively. Shading behind convergence measures shows the 95% CI of the uncertainty of the convergence measure. Grey line is converging proportional to $n^{-\frac{1}{2}}$. Single grid point data from the IFS ensemble used.

The convergence measure of the EFI for the 2m temperature as well as the total precipitation are calculated and shown in Figure 4.7 for the single grid point located in Figure 4.1. In line with how ECMWF computes the index, the one control member as well as the 50 ensemble members are used to create the model climate and ensemble distributions. The EFI is designed to highlight differences between the forecast and climatological distribution and as such, the forecast lead time of 3 hours is chosen for analysis. Comparing the magnitudes of the convergence measure is irrelevant as they are for different statistical quantities but what is interesting is that it looks as though both quantities are converging proportional to $n^{-\frac{1}{2}}$ with less than 10 members. This follows the convergence characteristics of the mean statistic.

As in Figure 4.6, the uncertainty of the bootstrap is calculated and shown by the background shading in Figure 4.7. For both variables, the uncertainty of the bootstrap is parallel to the convergence measure, indicating that the convergence measure gives a realistic estimate of the real convergence of sampling uncertainty with ensemble size. This is expected to be possible due to the underlying distribution shapes having no significant tails which could create a larger range of values in the nominator of Equation (18) and as such greater uncertainty in the value of the EFI. The confidence interval for the 2m temperature is less smooth than for the total precipitation. It is thought that this could be due to the temperature being relatively Gaussian symmetric whereas the total precipitation has a lot of density at zero, with only one relatively small tail to create uncertainty rather than two. This uncertainty analysis shows that the convergence measure of the EFI has converged asymptotically with just a 51-member ensemble.

From analysing the EFI, it has been seen that asymptotic convergence of sampling uncertainty is not limited to standard statistical quantities. This is theoretically possible since, as long as the statistic obeys the CLT, it will converge asymptotically with a large enough number of ensemble members. In the case of the EFI statistic, a 51-member ensemble was enough to observe this convergence. This suggests that the framework to estimate ensemble size which is based on asymptotic theory can also work with a great magnitude of commonly used operational forecast statistics as long as they obey the CLT.

4.4 Summary and Conclusions

A framework for estimating ensemble size in convective-scale data exists which is based on quantifying how sampling uncertainty converges with ensemble size. In the limit of large ensemble size convergence is proportional to $n^{-\frac{1}{2}}$, allowing one to estimate how many members are needed to reach a certain level of sampling uncertainty. In this chapter, data from a synoptic-scale forecast model was analysed in order to determine whether this framework to estimate ensemble size is not only applicable to convective scales but to synoptic scales in addition. Furthermore it is of interest whether the asymptotic convergence of sampling uncertainty applies not only to standard statistics but also to a common statistic used in operational forecasting, the EFI.

The 50-member synoptic-scale IFS ensemble forecast from ECMWF was first evaluated to determine if the distribution shapes were qualitatively similar to convective-scale ensembles. Since it is known that the underlying distribution shape significantly influences the convergence of sampling uncertainty, if similar shapes were found, this would likely mean that similar convergence behaviour would also be found. This was indeed the case as the three categories of distribution shape from convective-scale models were seen in the variables tested, which were the temperature at 2m and 500hPa, the relative humidity and the total precipitation. In addition, the evolution of the distributions in time followed a conceptual convective-scale framework.

With the 50-member ensemble, convergence of sampling uncertainty proportional to $n^{-\frac{1}{2}}$ was seen for the mean and most standard deviation quantities, as expected from convective-scale results. Larger ensemble sizes were needed for all quantiles tested however. The neighbourhood method was useful in expanding the effective ensemble size despite the more highly correlated nature of the synoptic-scale data. As long as the neighbourhood distributions were statistically similar and large enough, they could increase the effective ensemble size in order to reach asymptotic convergence for the variables tested.

The EFI statistical quantity was tested for asymptotic convergence as a quantity which is often employed operationally in the prediction of extreme events. Due to the EFI appearing to obey the CLT, the quantity was seen to converge proportional to $n^{-\frac{1}{2}}$ with less than 10 members for the 2m temperature and the total precipitation. This indicates that the sampling uncertainty of operational forecasting statistics are also able to converge asymptotically.

Asymptotic convergence of sampling uncertainty has been found to not only occur in the convective scale, but also in the synoptic scale. Furthermore, asymptotic convergence behaviour has been seen in statistics other than the standard moments of the distribution (e.g. mean and standard deviation) and quantiles but also for the EFI, a forecast statistic comparing the model climate and forecast distributions. The framework developed in Chapter 2 for estimating ensemble size can hence be of relevance to the synoptic scales.

Chapter 5

Conclusions

The purpose of this thesis was to answer the big question of **how to know the ensemble size required to achieve the desired accuracy in your statistic of interest**. This would allow for more precise weather forecast as well as for the optimal use of finite computational resources.

To find the required ensemble size, the approach of looking at how sampling uncertainty converges with ensemble size was taken. This followed on from our previous study (Craig et al., 2022) where we highlighted the potential of using asymptotic theory to create a framework to find the required ensemble size. By using idealised representations of the atmosphere, huge ensembles could be created with which the convergence could be investigated out to very large ensemble sizes - a task which is not possible with current operational forecasting models.

How sampling uncertainty decreases with ensemble size was first explored. The idealised model of Wuersch and Craig (2014) was developed to create an ensemble with $100,000$ members which was sufficiently representative of a convective-scale ensemble. That is, it satisfied the space and time scales of convection, it could model non-linear processes and the forecast distributions fit into the three categories of Gaussian, multimodal and highly skewed. Using a method which was developed using bootstrapping, 95% CIs of the sampling uncertainty, known as the convergence measure, could be plotted as a function of ensemble size, allowing one to visualise how sampling uncertainty converged with ensemble size.

An asymptotic power law scaling proportional to $n^{-\frac{1}{2}}$, where n is ensemble size, was observed in the convergence measure. This was seen universally in the limit of large ensemble size across all model variables and statistics tested and occurred as a result of the CLT. In extreme statistics such as a quantile of a distribution where there was not significant probability density at that quantile level, more ensemble members were required for convergence proportional to $n^{-\frac{1}{2}}$ to be observed. In comparison, fewer than 10 members were necessary for the mean and around 50 for the variance, which would be possible to be detected by today's operational ensemble sizes. The magnitude of the sampling uncertainty was additionally seen to depend on the shape of the distribution.

Two methods were developed to determine whether one was in the asymptotic convergence regime and what to do if one was not. The first method, which measured the uncertainty of the convergence measure, could let one know how certain the estimate of sampling uncertainty convergence was. If one was converging proportional to $n^{-\frac{1}{2}}$ with high estimation accuracy, one could be confident one was in the asymptotic regime. If one was not, the second parametric method, which used the distinct shapes of the forecast variables could be employed to create a statistically equivalent ensemble, which could be as large as needed to reach the asymptotic regime. Asymptotic convergence of sampling uncertainty proportional to $n^{-\frac{1}{2}}$ then allowed for a framework to determine ensemble size. If the convergence measure was converging asymptotically, the convergence measure could be extrapolated to smaller sampling uncertainties. Once the desired level of sampling uncertainty was reached, the corresponding ensemble size would then be the size of ensemble needed.

The next question was **how does convergence of sampling uncertainty with ensemble size differ between the convective weather regimes of weak and strong forcing?** An extended version of the idealised ensemble was developed which allowed for weak and strong forcing regimes. It was found that in periods of precipitation, weak forcing regimes will have the largest sampling uncertainty in extreme quantiles of moisture variables due to the small densities in the tails of their distributions. This furthermore leads to larger ensemble sizes required to reach the asymptotic regime. In addition, large differences in spread during specific time periods between the weak and strong forcing regimes for all variables led to large differences in sampling uncertainty between regimes for the mean and standard deviation statistic. It has hence been established that the different distribution shapes in weak and strong forcing regimes lead to differences in the sampling uncertainty convergence behaviour. Assuming that other regimes will also have distribution shapes specific to that regime, it is likely that, dependent on the weather regime, different ensemble sizes will be required to reach the same level of accuracy in the statistic of interest and forecast variable.

What size of ensemble to have is also a pressing question on larger synoptic scales and as such the final question was **whether the nature of sampling uncertainty convergence is the same for both the convective scale and the synoptic scale?**. Using synoptic-scale data from a 50-member ECMWF ensemble forecast, the distributions of the temperature at 2m and 500hPa, the total precipitation as well as the relative humidity were found to contain the three types of distribution shapes common to convective-scale forecasting. This led to the sampling uncertainty convergence behaviour on the convective and synoptic scales having similar characteristics. The neighbourhood method was additionally found to be of benefit in expanding the effective ensemble size to reach asymptotic convergence in cases where the forecast variable distribution did not change substantially within the neighbourhood region. This analysis led to the conclusion that the framework to find ensemble size that was previously developed for the convective scale also has potential for the synoptic scale.

Using the framework to find ensemble size that was developed in this thesis has far-reaching consequences. From continually comparing the idealised ensembles with larger, more complex models and also working with the operational data from ECMWF, this ensured that the results from the idealised ensembles are applicable to the real-world atmosphere. By finding the optimal size for an ensemble on the convective or synoptic scale, which is tailored to the specific forecast situation and the forecaster's needs, one can optimise finite computational resources. Dependent on the statistic of interest, forecast variable and acceptable level of sampling uncertainty, this may lead to needing a larger or smaller ensemble for the forecaster's purposes. This allows the forecast ensemble to be built or improved with the optimisation of ensemble size as a priority so that the wanted precision is met, and ensures less waste of computational resources.

There are a few limitations with the results presented in this thesis. Through the framework proposed, it is possible to determine how many ensemble members would be required to limit the sampling uncertainty to within a desired level. Depending on how small this uncertainty is desired to be however, this ensemble size may not be physically possible - not just because of lack of computational power, but due to the number of observations required for the validation of the ensemble prediction system. This is investigated for the evaluation of the Discrete Ranked Probability Score (DRPS) statistic which measures how close the observations are to the ensemble output. For the decomposition of this statistic into useful components, Candille and Talagrand (2005) calculated that the number of possible outcomes from the ensemble must be small compared to the number of observations. It was mentioned however that a different decomposition method could be used which does not encounter this limitation. Another caveat is that the framework proposed considers only sampling uncertainty and its dependence on ensemble size. 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. A final caveat is that in Chapter 4, only one single forecast case initialised on the 10th June 2021 was analysed for a grid point in Central Germany. Variations of the results obtained would have been seen if different grid points, times, as well as neighbourhood sizes had been used, especially since the distributions of forecast variables at different forecast lead times from different neighbourhood sizes sometimes showed more differences amongst themselves than the neighbourhood example chosen. Furthermore, it could be that certain model variables are more uncorrelated between grid points than others, leading to the neighbourhood method being less, or more effective. For example, it could be expected that the two variables tested in this thesis, 2m temperature and total precipitation, that they would have less correlation between grid points than say the temperature at 500hPa. Chapter 4 is therefore meant as a preliminary study into the applicability of asymptotic theory for determining ensemble size on the synoptic scale.

This thesis provides significant progress towards understanding ensemble size, however there is scope for further investigation. In this thesis, uni-variate distributions were primarily considered as this was sufficient in order to detect and analyse the convergence behaviour of sampling uncertainty as ensemble size increased. It may be of interest, however, to carry out further investigations of convergence of sampling uncertainty with multivariate distributions, since they have important applications such as predicting road conditions in the winter by looking at joint probability distributions of temperature and precipitation (Berrocal et al., 2010) and predicting extreme rainfall by looking at joint spatial precipitation distributions (Debusho and Diriba, 2021). Furthermore, in this thesis, primarily the standard statistics of the mean, standard deviation/variance and quantiles were analysed, with the EFI as the only statistic specifically tailored to operational forecasting. As other operational forecast statistics are expected to follow the CLT (e.g. the CRPS (Zamo and Naveau, 2018)), it is expected that they will also converge asymptotically in the limit of large ensemble size. A final thought is the need for further studies which look at how computational resources are allocated in a forecasting system. By considering the forecast variables and statistics of interest in the allocation of computational resources, perhaps new priorities in the allocation would be discovered since in this thesis it has been shown that the ensemble size needed depends strongly on the forecast variable and the statistic of interest.

Chapter 6

Appendix

6.1 Data Used for Fitting Convergence Measure

Model variable	Distribution	Statistic	Fitting cut-off
wind	c of Figure 2.3	mean	0
		variance	100
		0.6 quantile	0
		0.7 quantile	0
		0.95 quantile	100
		0.99 quantile	200
		0.999 quantile	2000
	h of Figure 2.3	mean	0
		variance	100
height	c of Figure 2.4	mean	3
		variance	100
		0.3 quantile	500
		0.375 quantile	80000
		0.4 quantile	2000
		0.6 quantile	30
		0.999 quantile	1000
	h of Figure 2.4	mean	3
		variance	100
rain	c of Figure 2.5	mean	3
		variance	100
		0.6 quantile	5
		0.7 quantile	5
		0.95 quantile	100
		0.99 quantile	300
		0.999 quantile	70000
	h of Figure 2.5	mean	3
		variance	100

Table 6.1: Convergence measure data used for fitting of $y = an^{-\frac{1}{2}}$. Data up to a certain ensemble size cut-off (column 4) was not used in fitting procedure.

6.2 Further ECMWF Distributions

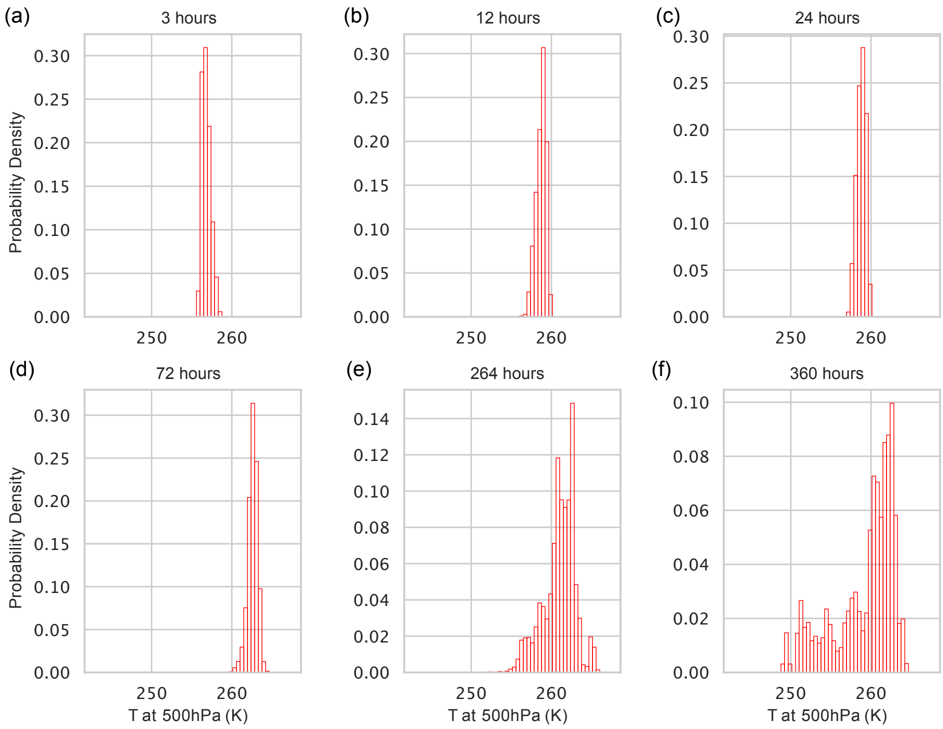

Figure B. 6.1: As in Figure 4.2 but for the temperature at 500hPa.

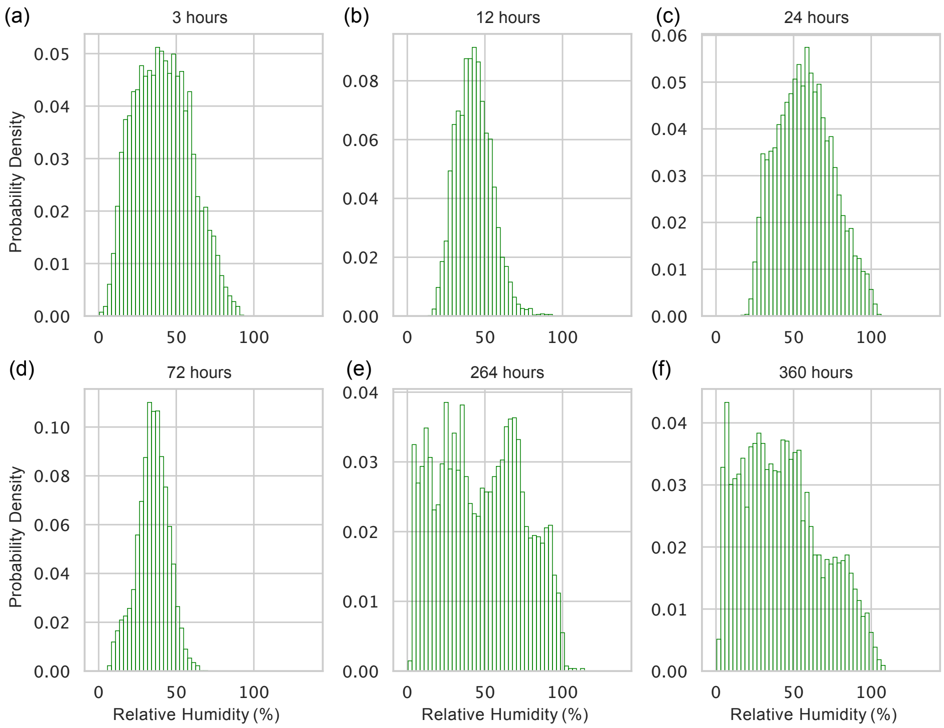

Figure B. 6.2: As in Figure 4.2 but for the relative humidity.

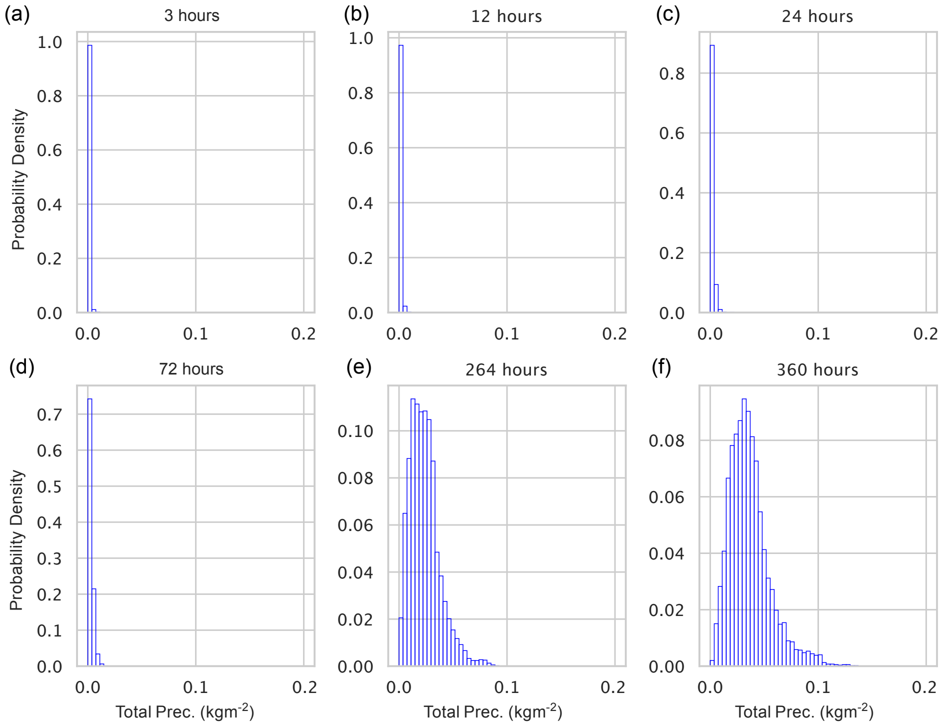

Figure B. 6.3: As in Figure 4.2 but for the total precipitation.

Acknowledgements

Firstly, I would like to thank my supervisor Prof. Dr. George Craig for guiding me through the last three and a half years. I'm lucky to have found such a brilliant scientist to work alongside as well as a patient and inspirational mentor, in the development of myself as a scientist and also in my career.

Although the Covid pandemic limited the time spent in office, I value the relationships made with colleagues. They provided friendly faces and a supporting hand. Special shout out to Mirjam Hirt, Tobias Selz and Matjaž Puh with whom I shared an office but also to the rest who were a stones throw away down the corridor. I enjoyed getting to know you, not only during the weekdays but also from hiking, road biking and skiing together amongst other things. I also enjoyed getting to know my colleagues in Waves to Weather and the meetings we had together throughout Germany, from which I am lucky to have made lasting friendships.

Thank you in addition to Dr. Christian Keil, my second supervisor as well as Scientific Manager Dr. Audine Laurian in supporting me throughout my doctoral studies.

This thesis is the accumulation not only of the last few years of work but rather of all the time I've been in education. The ones always standing by me, encouraging me and helping me where they could, were my parents. Whether that was helping me understand high-school physics homework problems or helping me to improve my writing skills. So a very special thank you goes to them.

Last but not least, thanks to my loving fiancé Florian. The years spent creating this thesis has been as much an academic journey as well as a journey of getting to know you. The Covid pandemic meant that we spent a significant portion of time together in our own home office, where you were my colleague. It was fun, and inspirational at points, talking with you about ensembles and getting your electrical engineer viewpoint on weather forecasting. Apart from that, I always had you to support me emotionally so that I was in a good position for completing this thesis.

Many of the above who were mentioned spent time helping me compile and refine this thesis. Thank you again for your efforts.

List of Abbreviations

MSE	Mean Square Error
EnKF	ensemble Kalman filter
NWP	Numerical Weather Prediction
CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
EFI	Extreme Forecast Index
DA	Data Assimilation
CI	Confidence Interval
CRPS	Continuous Ranked Probability Score
CLT	Central Limit Theorem
PDF	Probability Density Function
RMSE	Root Mean Square Error
KL Divergence	Kullback-Leibler Divergence
iid	independent and identically distributed
KDE	Kernel Density Estimation
DRPS	Discrete Ranked Probability Score
ECMWF	European Centre for Medium-range Weather Forecasts
EFI	Extreme Forecast Index
EDA	Ensemble of Data Assimilations
IFS	Integrated Forecast System
CAPE	Convective Available Potential Energy
SPPT	Stochastic Perturbation to Physical Tendencies
LNB	Level of Neutral Buoyancy
LFC	Level of Free Convection
CIN	Convective Inhibition

List of Figures

List of Tables

Bibliography

Abe, C., 1901: The physical basis of long-range weather forecasts. *Monthly Weather Review*, **29 (12)**, 551 – 561, doi:10.1175/1520-0493(1901)29[551c:TPBOLW]2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/29/12/1520-0493_1901_29_551c_tpbolw_2_0_co_2.xml.

Anderson, G., M. Kootval, H. Kootval, D. W. Kull, J. Clements, S. Consulting, G. Fleming, M. Éireann, T. Frei, J. K. Lazo, et al., 2017: Valuing weather and climate: Economic assessment of meteorological and hydrological services. *Disclosure*.

Bachmann, K., C. Keil, G. C. Craig, M. Weissmann, and C. A. Welzbacher, 2020: Predictability of deep convection in idealized and operational forecasts: Effects of radar data assimilation, orography, and synoptic weather regime. *Monthly Weather Review*, **148 (1)**, 63–81.

Bannister, R. N., S. Migliorini, A. C. Rudd, and L. H. Baker, 2017: Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble. *Geoscientific Model Development Discussions*, 1–38.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55, doi:10.1038/nature14956, URL https://doi.org/10.1038/nature14956.

Berrocal, V. J., A. E. Raftery, T. Gneiting, and R. C. Steed, 2010: Probabilistic weather forecasting for winter road maintenance. *Journal of the American Statistical Association*, **105 (490)**, 522–537.

Bjerknes, V., 1904: Das problem der wettervorhersage, betrachtet vom standpunkte der mechanik und der physik. *Meteor. Z.*, **21**, 1–7, URL https://cir.nii.ac.jp/crid/1570009750573730176.

Bolin, B., 1955: Numerical forecasting with the barotropic model1. *Tellus*, **7 (1)**, 27–49, doi:https://doi.org/10.1111/j.2153-3490.1955.tb01139.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1955.tb01139.x.

Buizza, R., J. Barkmeijer, T. N. Palmer, and D. S. Richardson, 2000: Current status and future developments of the ecmwf ensemble prediction system. *Meteorologi-*

cal Applications, **7 (2)**, 163–175, doi:https://doi.org/10.1017/S1350482700001456, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1017/S1350482700001456.

Buizza, R., T. Petroliagis, T. Palmer, J. Barkmeijer, M. Hamrud, A. Hollingsworth, A. Simmons, and N. Wedi, 1998: Impact of model resolution and ensemble size on the performance of an ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **124 (550)**, 1935–1960, doi:https://doi.org/10.1002/qj.49712455008, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712455008.

Candille, G. and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **131 (609)**, 2131–2150.

Charney, J., 1948: *On the scale of the atmospheric motions*. Geofys. Publ, London.

Cohen, B. G. and G. C. Craig, 2006: Fluctuations in an equilibrium convective ensemble. part ii: Numerical experiments. *Journal of the Atmospheric Sciences*, **63 (8)**, 2005 – 2015, doi:10.1175/JAS3710.1, URL https://journals.ametsoc.org/view/journals/atsc/63/8/jas3710.1.xml.

Craig, G. C. and B. G. Cohen, 2006: Fluctuations in an equilibrium convective ensemble. part i: Theoretical formulation. *Journal of the Atmospheric Sciences*, **63 (8)**, 1996 – 2004, doi:10.1175/JAS3709.1, URL https://journals.ametsoc.org/view/journals/atsc/63/8/jas3709.1.xml.

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

Craig, G. C., A. H. Fink, C. Hoose, T. Janjić, P. Knippertz, A. Laurian, S. Lerch, B. Mayer, A. Miltenberger, R. Redl, M. Riemer, K. I. Tempest, and V. Wirth, 2021: Waves to weather: Exploring the limits of predictability of weather. *Bulletin of the American Meteorological Society*, **102 (11)**, E2151 – E2164, doi:https://doi.org/10.1175/BAMS-D-20-0035.1, URL https://journals.ametsoc.org/view/journals/bams/102/11/BAMS-D-20-0035.1.xml.

Davison, A. and D. Hinkley, 1997: *Bootstrap Methods and their Applications*. Statistical and Probabilistic Mathematics, Cambridge University Press.

Debusho, L. K. and T. A. Diriba, 2021: Conditional modelling approach to multivariate extreme value distributions: application to extreme rainfall events in south africa. *Environmental and Ecological Statistics*, **28 (3)**, 469–501.

Dee, D. P., 2004: Variational bias correction of radiance data in the ecmwf system. *Proceedings of the ECMWF workshop on assimilation of high spectral resolution sounders in NWP, Reading, UK*, Vol. 28, 97–112.

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

Dibike, Y. B., P. Gachon, A. St-Hilaire, T. B. Ouarda, and V. T.-V. Nguyen, 2008: Uncertainty analysis of statistically downscaled temperature and precipitation regimes in northern canada. *Theoretical and Applied Climatology*, **91 (1)**, 149–170.

Done, J., G. Craig, S. Gray, P. A. Clark, and M. Gray, 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 (616)**, 737–756.

ECMWF, 2021a: *IFS Documentation CY47R3 - Part II: Data assimilation*. 2, ECMWF, doi:10.21957/t445u8kna, URL https://www.ecmwf.int/node/20196.

ECMWF, 2021b: *IFS Documentation CY47R3 - Part IV Physical processes*. 4, ECMWF, doi:10.21957/eyrpir4vj, URL https://www.ecmwf.int/node/20198.

ECMWF, 2021c: *IFS Documentation CY47R3 - Part V Ensemble prediction system*. 5, ECMWF, doi:10.21957/zw5j5zdz5, URL https://www.ecmwf.int/node/20199.

ECMWF, 2023: Extreme forecast index. ECMWF, URL https://confluence.ecmwf.int/display/FUG/Extreme+Forecast+Index+-+EFI2C+and+Shift+of+Tails+-+SOT.

Emanuel, K. A., 1955: *Atmospheric Convection*. Oxford University Press.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21 (6)**, 739–759, doi:10.3402/tellusa.v21i6.10143, URL https://doi.org/10.3402/tellusa.v21i6.10143, https://doi.org/10.3402/tellusa.v21i6.10143.

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 (C5)**, 10 143–10 162, doi:https://doi.org/10.1029/94JC00572, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572.

Feng, X., T. DelSole, and P. Houser, 2011: Bootstrap estimated seasonal potential predictability of global temperature and precipitation. *Geophysical Research Letters*, **38 (7)**.

Flack, D. L., R. S. Plant, S. L. Gray, H. W. Lean, C. Keil, and G. C. Craig, 2016: Characterisation of convective regimes over the british isles. *Quarterly Journal of the Royal Meteorological Society*, **142 (696)**, 1541–1553.

Gaspari, G. and S. E. Cohn, 1999: Construction of correlation functions in two and three dimensions. *Quarterly Journal of the Royal Meteorological Society*, **125 (554)**, 723–757, doi:https://doi.org/10.1002/qj.49712555417, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555417.

Gneiting, T., 2014: Calibration of medium-range weather forecasts. **(719)**, doi:10.21957/8xna7glta, URL https://www.ecmwf.int/node/9607.

Harding, B., C. Tremblay, and D. Cousineau, 2014: Standard errors: A review and evaluation of standard error estimators using monte carlo simulations. *The Quantitative Methods for Psychology*, **10 (2)**, 107–123.

Hendricks, E. A., J. L. Vigh, and C. M. Rozoff, 2021: Forced, balanced, axisymmetric shallow water model for understanding short-term tropical cyclone intensity and wind structure changes. *Atmosphere*, **12 (10)**, 1308.

Hirt, M., S. Rasp, U. Blahak, and G. C. Craig, 2019: Stochastic parameterization of processes leading to convective initiation in kilometer-scale models. *Monthly Weather Review*.

Isaksen, L., M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, and L. Raynaud, 2010: Ensemble of data assimilations at ecmwf.

Jacques, D. and I. Zawadzki, 2015: The impacts of representing the correlation of errors in radar data assimilation. part ii: Model output as background estimates. *Monthly Weather Review*, **143 (7)**, 2637 – 2656, doi:10.1175/MWR-D-14-00243.1, URL https://journals.ametsoc.org/view/journals/mwre/143/7/mwr-d-14-00243.1.xml.

Jirak, I. L., A. J. Clark, C. J. Melick, and S. J. Weiss, 2016: 15b. 5 investigation of the impact of convection-allowing ensemble size for severe weather forecasting.

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Weather and Forecasting*, **22 (3)**, 637–650.

Kawabata, T. and G. Ueno, 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 (1)**, 3–20.

Keil, C., F. Baur, K. Bachmann, S. Rasp, L. Schneider, and C. Barthlott, 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 (724)**, 3102–3115.

Keil, C. and G. C. Craig, 2011: Regime-dependent forecast uncertainty of convective precipitation. *Meteorologische Zeitschrift*, **20 (2)**, 145.

Keil, C., F. Heinlein, and G. C. Craig, 2014: The convective adjustment time-scale as indicator of predictability of convective precipitation. *Quarterly Journal of the Royal Meteorological Society*, **140 (679)**, 480–490.

Kondo, K. and T. Miyoshi, 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 (3)**, 211–225, doi:10.5194/npg-26-211-2019, URL https://npg.copernicus.org/articles/26/211/2019/.

Kullback, S. and R. A. Leibler, 1951: On information and sufficiency. *The Annals of Mathematical Statistics*, **22 (1)**, 79–86, URL http://www.jstor.org/stable/2236703.

Kuo, Y.-H. and R. J. Reed, 1988: Numerical simulation of an explosively deepening cyclone in the eastern pacific. *Monthly Weather Review*, **116 (10)**, 2081–2105.

Lakatos, M., S. Lerch, S. Hemri, and S. Baran, 2022: Comparison of multivariate post-processing methods using global ecmwf ensemble forecasts. 2206.10237.

Lang, S., E. Hólm, M. Bonavita, and Y. Tremolet, 2019: A 50-member ensemble of data assimilations. URL https://www.ecmwf.int/node/18883, 27-29 pp., doi:10.21957/nb251xc4sl.

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

Leith, C. E., 1974: Theoretical skill of monte carlo forecasts. *Monthly weather review*, **102 (6)**, 409–418.

Leutbecher, M., 2019: Ensemble size: How suboptimal is less than infinity? *Quarterly Journal of the Royal Meteorological Society*, **145 (S1)**, 107–128, doi:https://doi.org/10.1002/qj.3387, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3387.

Leutbecher, M. and S. Lang, 2014: On the reliability of ensemble variance in subspaces defined by singular vectors. *Quarterly Journal of the Royal Meteorological Society*, **140 (682)**, 1453–1466.

Leutbecher, M. and T. N. Palmer, 2008: Ensemble forecasting. *Journal of computational physics*, **227 (7)**, 3515–3539.

Lin, J., K. Emanuel, and J. Vigh, 2020: Forecasts of hurricanes using large-ensemble outputs. *Weather and Forecasting*, **5 (35)**, 1713–1731, doi:10.1175/WAF-D-19-0255.1.

Lin, Y.-L., 2007: *Mesoscale dynamics*. Cambridge University Press.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *Journal of atmospheric sciences*, **20 (2)**, 130–141.

Lorenz, E. N., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus*, **17 (3)**, 321–333.

Lovejoy, S. and D. Schertzer, 2018: *The weather and climate: emergent laws and multifractal cascades*. Cambridge University Press.

Machete, R. L. and L. A. Smith, 2016: Demonstrating the value of larger ensembles in forecasting physical systems. *Tellus A: Dynamic Meteorology and Oceanography*, **68 (1)**, 28 393.

Metoffice, 2023: The met office ensemble system. Metoffice, URL https://www.metoffice.gov.uk/research/weather/ensemble-forecasting/mogreps.

Milinski, S., N. Maher, and D. Olonscheck, 2020: How large does a large ensemble need to be? *Earth System Dynamics*, **11 (4)**, 885–901.

Miyoshi, T., K. Kondo, and T. Imamura, 2014: The 10,240-member ensemble kalman filtering with an intermediate agcm. *Geophysical Research Letters*, **41 (14)**, 5264–5271, doi:https://doi.org/10.1002/2014GL060863, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014GL060863.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ecmwf ensemble prediction system: Methodology and validation. *Quarterly journal of the royal meteorological society*, **122 (529)**, 73–119.

Molteni, F. and T. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quarterly Journal of the Royal Meteorological Society*, **119 (510)**, 269–298.

Necker, T., S. Geiss, M. Weissmann, J. Ruiz, T. Miyoshi, and G.-Y. Lien, 2020a: A convective-scale 1,000-member ensemble simulation and potential applications. *Quarterly Journal of the Royal Meteorological Society*, **146 (728)**, 1423–1442.

Necker, T., M. Weissmann, Y. Ruckstuhl, J. Anderson, and T. Miyoshi, 2020b: Sampling error correction evaluated using a convective-scale 1000-member ensemble. *Monthly Weather Review*, **148 (3)**, 1229 – 1249, doi:10.1175/MWR-D-19-0154.1, URL https://journals.ametsoc.org/view/journals/mwre/148/3/mwr-d-19-0154.1.xml.

Palmer, T., 2017: The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, **9 (2)**, 730–734.

Palmer, T., 2019: The ecmwf ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, **145**, 12–24.

Palmer, T., F. Molteni, R. Mureau, R. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. Ph.D. thesis, 21-66 pp., Shinfield Park, Reading.

Petrie, R. E., R. N. Bannister, and M. J. P. Cullen, 2017: The abc model: a non-hydrostatic toy model for use in convective-scale data assimilation investigations. *Geoscientific Model Development*, **10 (12)**, 4419–4441.

Poterjoy, J., 2022: Implications of multivariate non-gaussian data assimilation for multiscale weather prediction. *Monthly Weather Review*, **150 (6)**, 1475 – 1493, doi:10.1175/MWR-D-21-0228.1, URL https://journals.ametsoc.org/view/journals/mwre/150/6/MWR-D-21-0228.1.xml.

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

Raynaud, L. and F. Bouttier, 2017: The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, **143 (709)**, 3037–3047, doi:https://doi.org/10.1002/qj.3159, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3159.

Reinert, D., F. Prill, H. Frank, M. Denhard, M. Baldauf, C. Schraff, C. Gebhardt, C. Marsigli, and G. Zängl, 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 (577)**, 2473–2489, doi:https://doi.org/10.1002/qj.49712757715, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712757715.

Rodwell, M. J. and H. Wernli, 2022: The cyclogenesis butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis. *Weather and Climate Dynamics Discussions*, **2022**, 1–32, doi:10.5194/wcd-2022-6, URL https://wcd.copernicus.org/preprints/wcd-2022-6/.

Ruckstuhl, Y., T. Janjić, and S. Rasp, 2021: Training a convolutional neural network to conserve mass in data assimilation. *Nonlinear Processes in Geophysics*, **28 (1)**, 111–119, doi:10.5194/npg-28-111-2021, URL https://npg.copernicus.org/articles/28/111/2021/.

Ruckstuhl, Y. M. and T. Janjić, 2018: Parameter and state estimation with ensemble kalman filter based algorithms for convective-scale applications. *Quarterly Journal of the Royal Meteorological Society*, **144 (712)**, 826–841.

Scheuerer, M. and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, **143 (11)**, 4578–4596.

Scott, D. W., 2015: *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons.

Selz, T., M. Riemer, and G. C. Craig, 2022: The transition from practical to intrinsic predictability of midlatitude weather. *Journal of the Atmospheric Sciences*, **79 (8)**, 2013–2030.

Slingo, J. and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **369 (1956)**, 4751–4767.

Stuart, A. and J. Ord, 2000: *Kendall's Advanced Theory of Statistics*, Vol. 1. Arnold Publishers.

Tempest, K. I., G. C. Craig, and J. R. Brehmer, 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, doi:https://doi.org/10.1002/qj.4410, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.4410.

Thomas, S., P. Prabhakaran, W. Cantrell, and R. A. Shaw, 2021: Is the water vapor supersaturation distribution gaussian? *Journal of the Atmospheric Sciences*, **78 (8)**, 2385–2395.

Thomson, P. D., 1957: Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus*, **9 (3)**, 275–295, doi:https://doi.org/10.1111/j.2153-3490.1957.tb01885.x, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1957.tb01885.x.

Tompkins, A., 2005: The parameterization of cloud cover.

Tompkins, A. M. and M. Janisková, 2004: A cloud scheme for data assimilation: Description and initial tests. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, **130 (602)**, 2495–2517.

Urdan, T. C., 2022: *Statistics in plain English.* Taylor & Francis.

Wang, P., J. Li, M. D. Goldberg, T. J. Schmit, A. H. N. Lim, Z. Li, H. Han, J. Li, and S. A. Ackerman, 2015: Assimilation of thermodynamic information from advanced infrared sounders under partially cloudy skies for regional nwp. *Journal of Geophysical Research: Atmospheres*, **120 (11)**, 5469–5484, doi:https://doi.org/10.1002/2014JD022976, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014JD022976.

WEF, 2020: The global risks report 2020. *Davos: World Economic Forum. Retrieved November*, Vol. 15, 2020.

Williams, P. D., 2009: A proposed modification to the robert–asselin time filter. *Monthly Weather Review*, **137 (8)**, 2538 – 2546, doi:10.1175/2009MWR2724.1, 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 (6)**, 1996 – 2007, doi:10.1175/2010MWR3601.1, URL https://journals.ametsoc.org/view/journals/mwre/139/6/2010mwr3601.1.xml.

Wuersch, M. and G. C. Craig, 2014: A simple dynamical model of cumulus convection for data assimilation research. *Meteorologische Zeitschrift*, **23 (5)**, 483–490, doi:10.1127/0941-2948/2014/0492, URL http://dx.doi.org/10.1127/0941-2948/2014/0492.

Yang, D., 2021: A shallow-water model for convective self-aggregation. *Journal of the Atmospheric Sciences*, **78 (2)**, 571–582.

Zamo, M. and P. Naveau, 2018: Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, **50 (2)**, 209–234.

Zhang, F., 2005: Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts. *Monthly Weather Review*, **133 (10)**, 2876 – 2893, doi:10.1175/MWR3009.1, URL https://journals.ametsoc.org/view/journals/mwre/133/10/mwr3009.1.xml.

Zimmer, M., G. Craig, C. Keil, and H. Wernli, 2011: Classification of precipitation events with a convective response timescale and their forecasting characteristics. *Geophysical Research Letters*, **38 (5)**.