The Transition from Practical to Intrinsic Predictability of Midlatitude Weather

TOBIAS SELZ,a MICHAEL RIEMER,b AND GEORGE C. CRAIGa

a Meteorologisches Institut, Ludwig-Maximilians-Universität, Munich, Germany
b Institut für Physik der Atmosphäre, Johannes Gutenberg-Universität Mainz, Mainz, Germany

(Manuscript received 8 October 2021, in final form 28 February 2022)

ABSTRACT: This study investigates the transition from current practical predictability of midlatitude weather to its intrinsic limit. For this purpose, estimates of the current initial condition uncertainty of 12 real cases are reduced in several steps from 100% to 0.1% and propagated in time with a global numerical weather prediction model (ICON at 40 km resolution) that is extended by a stochastic convection scheme to better represent error growth from unresolved motions. With the provision that the perfect model assumption is sufficiently valid, it is found that the potential forecast improvement that could be obtained by perfecting the initial conditions is 4–5 days. This improvement is essentially achieved with an initial condition uncertainty reduction by 90% relative to current conditions, at which point the dominant error growth mechanism changes: With respect to physical processes, a transition occurs from rotationally driven initial error growth to error growth dominated by latent heat release in convection and due to the divergent component of the flow. With respect to spatial scales, a transition from large-scale up-amplitude error growth to a very rapid initial error growth on small scales is found. Reference experiments with a deterministic convection scheme show a 5%–10% longer predictability, but only if the initial condition uncertainty is small. These results confirm that planetary-scale predictability is intrinsically limited by rapid error growth due to latent heat release in clouds through an upscale-interaction process, while this interaction process is unimportant on average for current levels of initial condition uncertainty.

SIGNIFICANCE STATEMENT: Weather predictions provide high socioeconomic value and have been greatly improved over the last decades. However, it is widely believed that there is an intrinsic limit to how far into the future the weather can be predicted. Using numerical simulations with an innovative representation of convection, we are able to confirm the existence of this limit and to demonstrate which physical processes are responsible. We further provide quantitative estimates for the limit and the remaining improvement potential. These results make clear that our current weather prediction capabilities are not yet maxed out and could still be significantly improved with advancements in atmospheric observation and simulation technology in the upcoming decades.

KEYWORDS: Uncertainty; Ensembles; Numerical weather prediction/forecasting

1. Introduction

Despite great progress in recent decades, the practical predictability of midlatitude weather is currently limited to about 1 week for most users of the forecasts. While current forecast skill is restricted by significant errors and uncertainties in the initial conditions and the numerical model, it has now been established that predictability remains finite even if those uncertainties and errors were reduced toward zero. This “intrinsic limit of predictability” arises from scale interactions in a multiscale fluid system like the atmosphere and is not just a consequence of a high sensitivity to initial conditions, which is already present in low-order chaotic systems (Lorenz 1969; Palmer et al. 2014).

Two basic methods can be distinguished that have been used in the past to estimate the intrinsic limit. The first approach considers highly simplified statistical models based on two-dimensional homogeneous turbulence (e.g., Lorenz 1969; Rotunno and Snyder 2008; Durran and Gingrich 2014; Leung et al. 2020) to describe interactions between errors on different scales and in turn error growth rates and predictability times. It has been found that the most important parameter in determining predictability in such models is the slope of the kinetic energy spectrum (Rotunno and Snyder 2008). The second method takes comprehensive numerical weather prediction models and analyzes the time evolution of the difference between two or more simulations that differ initially by only a very small amount (e.g., Judt 2018, 2020; Selz 2019; Zhang et al. 2019).

Both types of studies agree that the order of magnitude of the intrinsic limit is about 2 weeks. However, the exact number

Denotes content that is immediately available upon publication as open access.

Corresponding author: Tobias Selz, tobias.selz@lmu.de

DOI: 10.1175/JAS-D-21-0271.1

© 2022 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).
depends on details of the definition of the limit such as the variable being predicted, the height level, the spatial and temporal scales, and the metric used to quantify predictability. Furthermore, the loss of predictability happens gradually and the specification of a predictable time interval requires the definition of some kind of threshold at which a forecast is no longer considered useful. The threshold depends on the demands of a certain user, but is often also dictated by the available sample size of the experiments. Large samples are able to detect small remnants of predictability, as demonstrated by Buizza and Leutbecher (2015), who found very weak but significant predictive skill up to 23 days in the current operational ECMWF ensemble forecasting system which is longer than the 2-week estimate for the intrinsic limit. These differences in metrics, thresholds, and sample sizes can make it difficult to compare studies on practical and intrinsic predictability and to estimate the remaining improvement potential.

Since laboratory experiments with a twin Earth cannot be done, every estimate of the intrinsic limit is model based and only valid to the extent that the approximations and simplifications of the model do not crucially affect the growth and propagation of the initial uncertainty (perfect model assumption). While simplified turbulence models ignore heterogeneity, three-dimensional dynamics, and diabatic processes, the main limitation of experiments with complex numerical models is their rather crude spatial resolution: The potentially important transition to a flatter kinetic energy spectrum at around the 400 km scale is simulated poorly, if at all, in current global models with $O(10-20)$ km resolution. Furthermore, moist processes and especially convection have been shown to play a major role in error growth if the initial condition uncertainty is small (Zhang et al. 2003; Hohenegger and Schär 2007; Selz and Craig 2015b). Convection is however to a large degree parameterized in global simulations, leading to errors. By directly comparing different spatial resolutions, Selz and Craig (2015a) showed that upscale error growth from convection is underestimated in simulations with $\geq 5$ km grid sizes and parameterized convection compared to convection-permitting simulations with 2.8 km grid size. Consequently, numerical experiments to determine the intrinsic limit that are based on weather prediction models should ideally be conducted at a convection-permitting resolution, which is computationally very challenging. This has however recently become feasible and such an experiment has been done with one pair of simulations for a single case (Judit 2018, 2020).

Selz and Craig (2015a) have further demonstrated that a stochastic convection scheme that reintroduces missing variability at the model’s grid scale could compensate for too slow upscale error growth from convection in lower-resolution simulations. Relying on this finding, Selz (2019) estimated the intrinsic limit with global simulations that used stochastic convection. The order-of-magnitude value of 2 weeks for the intrinsic limit was confirmed, but a direct comparison to simulations with deterministic convection showed that the stochastic convection produced an estimate that was shorter by about 10%.

In a subsequent analysis of the simulations of Selz (2019) by Baumgart et al. (2019), the physical processes responsible for error growth were analyzed by means of potential vorticity (PV) diagnostics. These diagnostics clearly identified an initial growth phase where PV error growth was mainly due to the stochastic convection scheme, followed by a transition phase where advection of PV by the divergent component of the flow was the main source of error, and then a phase of error growth associated with the rotational wind component, which lasted until the errors saturated. This picture is broadly consistent with the conceptual three-stage error-growth model proposed by Zhang et al. (2007), although the interpretation of the dominant mechanisms at the individual stages differ between these studies. In particular, tropospheric deep interactions (i.e., baroclinic instability) were found to be unimportant for error growth on average.

To identify when and how the intrinsic limit might impact the skill of future forecasting systems, it is necessary to understand the transition in error growth processes as initial condition accuracy improves. Sun and Zhang (2016) investigated predictability by means of idealized baroclinic wave experiments with a variety of initial condition perturbations. They found that for small-amplitude initial condition perturbations, rapid error growth on convective scales is triggered, which is insensitive to amplitude and scale of those perturbations. They identified this behavior with the intrinsic predictability limit. On the other hand, if the initial condition perturbations are large, error growth is sensitive to the amplitude and scale of the initial condition error, but insensitive to moist processes, which they argue represents the situation in current operational forecasts.

A direct comparison of the current practical and intrinsic predictability limits was conducted by Zhang et al. (2019), who used the ECMWF’s ensemble of data assimilations (EDA) system to quantify the current initial condition uncertainty. Ensemble simulations from the unchanged initial condition sample were compared to simulations from a sample where the spread was reduced to 10% of its original magnitude, which was assumed to represent the intrinsic limit. The study was based on six cases (3 consecutive days in summer and in winter) and it was found that this 90% decrease in initial condition uncertainty led to an increase in the predictability limit of about 5 days.

In the present study, we seek a deeper understanding of the transition from current practical predictability to intrinsic predictability of midlatitude weather by addressing the question of how far away the intrinsic limit is for current forecasting systems on average, what level of improvement would be required to reach it, and what the key characteristics are that distinguish the practical and the intrinsic predictability regime. To do this, we use a complex numerical weather prediction model where we approximate unresolved convective motions and related error growth with a stochastic convection scheme (as in Selz 2019). We follow Zhang et al. (2019) and reduce the current initial condition uncertainty, as estimated by the ECMWF EDA system, but in multiple steps from 100% down to 0.1%. To get representative results, we conduct global, 5 member ensemble simulations of 12 cases, distributed over one year. For this study we focus on midlatitude predictability using instantaneous, gridpoint metrics based on kinetic energy and potential vorticity at the tropopause level. Additional simulations with a standard deterministic convection scheme, with singular vectors added to
the initial conditions, and with twice the resolution are also included for reference and comparison. We complement common spectral error growth analyses with the potential vorticity diagnostics of Baumgart et al. (2019) to quantify the relative importance of different physical processes for error growth and their dependence on the amplitude of initial condition uncertainty.

2. Methods

a. Numerical model and stochastic convection scheme

The simulations for this study are conducted with the Icosahedral Nonhydrostatic Model (ICON; Zängl et al. 2015), a comprehensive numerical weather prediction model that has been developed at the German Weather Service (DWD) and the Max Planck Institute for Meteorology in Germany. ICON consists of a nonhydrostatic dynamical core on an icosahedral-triangular grid, with explicit time integration except for vertically propagating sound waves. For efficiency, time splitting is applied outside of the dynamical core by computing tracer advection and tendencies from parameterization schemes on longer time steps. For this study, no ocean model is included and the sea surface temperatures are fixed to the initial condition values. The land surface is modeled with TERRA (Heise et al. 2006), a multilayer soil and vegetation scheme with a tile-based approach to account for subgrid-scale land-use variability. It is complemented with a snow-cover, lake, and sea ice model. For gridscale precipitation (microphysics) a one-moment scheme based on Seifert (2008) is used.

The standard parameterization scheme for convection is a version of the Tiedtke–Bechtold scheme, adapted from ECMWF (Tiedtke 1989; Bechtold et al. 2008, 2014). This is a conventional mass-flux scheme that computes convective tendencies of temperature and other quantities deterministically from the gridscale variables in the column. These tendencies represent the average effect of a large ensemble of clouds within a grid box, and do not represent the variability that occurs if the gridbox size and hence the number of clouds per grid box are small.

To account for the missing variability and its potential effects on error growth, most of the simulations in this paper will use the stochastic convection scheme of Plant and Craig (2008). This scheme triggers a random number of convective plumes in a grid box with a probability that scales with the closure and with the gridbox size. This ensures that in the limit of large grid boxes (low resolution), many clouds per grid box are triggered and the resulting convective tendencies converge to those of a deterministic mass flux scheme. As the gridbox size (and hence the number of clouds) gets smaller the scheme generates more random variability. This variability is uncorrelated in space and does not account for subgrid-scale organization, but it follows the spatial envelope determined by the mass-flux closure. A temporal correlation is imposed by applying a fixed cloud lifetime, which is set to 48 min and reduces unrealistic high-frequency variability from the (deterministic) trigger scheme. Through this mechanism, the Plant–Craig convection scheme is designed to adapt to a wide range of spatial resolutions, which has been confirmed in numerical simulations (Keane et al. 2014; Selz and Craig 2015a).

In its current implementation, the scheme applies the Kain–Fritsch plume model (e.g., Kain 2004), which computes tendencies of heat, moisture, and hydrometeors, but does not simulate the vertical transport of horizontal momentum. The Plant–Craig scheme also only considers deep convection and therefore we run the deterministic Tiedtke–Bechtold scheme in shallow convection mode in addition, which leads to a significantly improved forecast quality compared to simulations without a shallow convection scheme.

b. Experimental design

The initial conditions for the ICON simulations in this study are taken from the ECMWF’s EDA system [see Isaksen et al. (2010) or the documentation at www.ecmwf.int]. The EDA system uses perturbed observations and a model uncertainty representation in the assimilation cycle to generate an ensemble of 50 initial conditions that sample the current uncertainties of the observational network and the assimilation process. To investigate the transition from practical to intrinsic predictability, we rescale the perturbations in this initial condition ensemble (the spread) by five different factors (100%, 50%, 20%, 10%, 0.1%), where 100% means no change to the EDA sample and 0.1% scales the initial condition spread down to a very tiny amount to represent the intrinsic limit. The perturbations to all variables are rescaled with the same factor and the spatial structure and hence the spectral composition of the perturbations is left unchanged. Zhang et al. (2019) used a similar approach, but considered only a single factor (10%).

For every EDA rescaling factor, 12 different cases are simulated, evenly distributed over 1 year (as in Selz 2019). The initialization dates of the cases are 1 October 2016, 1 November 2016, and so on until 1 September 2017. For each case, a subset of 5 members is chosen randomly from the 50-member EDA ensemble to save computational cost and data volume. This 5-member ensemble is integrated to 31 days’ lead time with the ICON model at a resolution of R2B6, which equals approximately 40 km and with 90 vertical levels and a model top at 75 km. In addition to the different initial conditions, every member of the ensemble uses a different random seed for the stochastic convection scheme, resulting in a different distribution of the convective clouds to sample the rapid unresolved convective-scale error growth. Note that unlike common model uncertainty representation schemes (e.g., SPPT; Buizza et al. 1999) the stochastic convection scheme in our study is not an attempt to represent model uncertainty that arises from various approximations and uncertainties in the design of the scheme, but to sample a physical variability that exists within a single model formulation and is not deterministically related to the resolved atmospheric state. It is therefore an attempt to reduce the model error with respect to upscale growth from the initial condition uncertainty due to convection and to make the perfect model assumption more accurate (Selz and Craig 2015a).
In the reminder of this paper, we will refer to all simulations that employ a certain EDA rescale factor as one experiment, which we will label as 100%S, 50%S, 20%S, 10%S, and 0.1%S, where the “S” indicates that the stochastic convection scheme was used to represent deep convection. As a reference, we further consider three additional experiments, 100%D, 10%D, and 0.1%D, where the deterministic Tiedtke–Bechtold convection scheme is used in the model for both deep and shallow convection. For those simulations, no random seed is needed and the model is deterministic.

Although the ECMWF forecasting system samples the initial condition uncertainty with the EDA and includes SPPT to sample model uncertainty, additional ensemble inflation is still required to improve reliability, particularly in the short term (Isaksson et al. 2010). This is achieved by adding singular vectors to the initial conditions, which are a set of perturbations that experience fastest growth over a specified forecast time interval (2 days in the ECMWF system). They are derived from a linearized model in terms of a singular value decomposition [see, e.g., Leutbecher and Palmer (2008) and Diaconescu and Laprise (2012) for an overview]. It is largely unknown, however, what shortcoming of the forecasting system the singular vectors account for. The initial condition uncertainty might be underrepresented by the EDA system or the model uncertainty may be underrepresented by SPPT. Palmer (2019) hypothesized that the singular vectors could account for too slow error growth that arises from insufficient resolution and representation of convection. The study of Zhang et al. (2019) did not include singular vectors in the initial conditions nor discuss their relevance, which has been criticized (Zagar and Szunyogh 2020). Nevertheless, we also decided to not include the singular vectors in the rescaled experiments. Since it is unclear to what extent they compensate missing initial condition uncertainty, it is even less clear if and how they should be rescaled. To quantify their effects in the context of current forecasting systems, however, we added one additional reference experiment (labeled 100%Ssv) where the unchanged EDA sample is used and the singular vector perturbations are included.

To create a reasonable statistical basis (i.e., several cases, members, rescale factors), we had to limit the horizontal resolution of our experiments to about 40 km, which is rather crude compared to current operational standards. Hence we repeated two of the stochastic convection experiments (100%S and 0.1%S) with twice the resolution (R2B7, approximately 20 km) to provide a basic test on resolution sensitivity. Due to the much higher computational cost, these experiments were only run for 12 and 15 days, respectively. We refer to them as 100%Shr and 0.1%Shr.

In summary, our study considers 11 different experiments (100%S, 50%S, 20%S, 10%S, 0.1%S, 100%D, 10%D, 0.1%D, 100%Ssv, 100%Shr, 0.1%Shr). Each experiment consists of the same 12 cases, simulated with a 5 member ensemble, giving 720 single ICON simulations in total (540 simulations at 40 km resolution with 31 day runtime and 120 simulations at 20 km resolution with 12 and 15 day runtime, respectively). Independent of the model resolution, data are output hourly and interpolated to a regular 1° spatial grid.

c. Diagnostics

1) DIFFERENCE KINETIC ENERGY

In many error growth and predictability studies (e.g., Lorenz 1969; Rotunno and Snyder 2008; Judt 2020; Leung et al. 2020), difference kinetic energy (DKE) has been used as metric. It has the advantage that it relates to the (background) kinetic energy (KE) of the atmosphere, which has been extensively studied in the context of predictability. KE and DKE per unit mass are defined by:

\[ e(x, y) = \frac{1}{2}(u^2 + v^2) \]

\[ \Delta e(x, y) = \frac{1}{2}[(\Delta u)^2 + (\Delta v)^2] \]

where \( u \) and \( v \) are zonal and meridional wind components, respectively, and \( \Delta u \) and \( \Delta v \) are the differences between two simulations. Note that the DKE is not the difference of the kinetic energy, but rather the kinetic energy of the difference wind and as such positive definite. For this study we select the 300 hPa level as representative of the tropopause region.

Since our experiments consist of five ensemble members and not just two simulations, we generalize the definition of DKE by taking the mean over every possible combination of pairs, which can be rewritten as

\[ \Delta e(x, y) = \text{var}(u) + \text{var}(v), \]

with the unbiased sample variance. The DKE can also be interpreted as the mean square error (MSE) of \( u \) and \( v \) with respect to an analysis, if the bias component is ignored (which is intended in a perfect model scenario). It can be related to probabilistic verification metrics like the continuous ranked probability score under certain conditions (Leung et al. 2021).

Predictability time estimates based on DKE can be computed by comparing it to the variance of a climatological reference. We define this reference taking 20 years (2000–19) of daily 0000 UTC wind data from the ERA5 reanalysis (Hersbach et al. 2020). The 20 years and the 31 days for each case make up a 620-member ensemble of climatological wind variability from which the climatological DKE (\( \Delta E_{c,lim} \)) is computed using (2) for each case. The DKE from the experiments and from the climatology are spatially averaged over the mid-latitudes (40°–60°) in either hemisphere and the predictability time is defined as the earliest forecast lead time \( t \), for which the following condition is fulfilled:

\[ \Delta E_{c,h}(t) \geq \alpha \Delta E_{c,lim} \]

where \( \Delta E \) indicates the spatial average of \( \Delta e \), \( c \) indicates the case (1–12), \( h \) indicates the hemisphere (S or N), and \( \alpha < 1 \) is a threshold to be chosen. For every experiment, this will result in 24 predictability times, from which a mean predictability time per experiment can be computed together with a 95% confidence interval derived from the variability.

In addition, a decomposition of KE and DKE into spatial scales can be computed with the help of a Fourier transformation. Since for this paper, we are only interested in midlatitude...
error growth, we will use one-dimensional Fourier transforms in the (periodic) longitudinal direction and assume constant zonal grid spacing, which makes it later possible to average over the midlatitude band. Hence, for a certain latitude, the spectral density of the kinetic energy is given by

\[
\hat{\varepsilon}(k) = k \frac{1}{2} (\langle |\tilde{u}_i|^2 \rangle + \langle |\tilde{v}_i|^2 \rangle),
\]

(4)

where \( k \) is the zonal wavenumber and \( \tilde{u}_i \) and \( \tilde{v}_i \) are the 1D Fourier transforms in the longitudinal direction of the zonal and meridional wind components of ensemble member \( i \). The angle brackets \( \langle \cdot \rangle \) denote ensemble averages. The spectral DKE for an ensemble can again be defined as the DKE averaged over all pairs, which can be written as

\[
\Delta\hat{\varepsilon} = k \frac{N}{N-1} (\langle |\tilde{u}_i|^2 \rangle + \langle |\tilde{v}_i|^2 \rangle - |\langle \tilde{u}_i \rangle|^2 - |\langle \tilde{v}_i \rangle|^2),
\]

(5)

where \( N = 5 \) is the ensemble size. In essence, the equation states that the DKE spectral density of an ensemble is given by the ensemble mean of the KE spectral densities minus the KE spectral density of the ensemble mean. The spectral DKE will converge toward twice the background KE if all predictability in a certain mode is lost, i.e., the phases of the mode in the ensemble are randomly distributed.

We require that the spectral densities of KE and DKE are normalized to satisfy the following form of the Parseval identity,

\[
\mathcal{E} := \frac{1}{A} \int dA \, e(x,y) = \int d\log k \, \hat{\varepsilon}(k,y),
\]

(6)

and analogously for DKE. The overbar here denotes the latitude average over 40°-60°N/S. Note that our definition of the spectral densities (4) and (5) follows Lorenz (1969), and deviates from the standard definition [see, e.g., Durran et al. 2017, Eq. (3)] by an additional \( k \) factor, since the spectral density is here defined per unit log \( k \), as expressed by (6), and not per linear \( k \). This modification has been made to represent the fact that different atmospheric scale ranges are usually separated by a constant multiplicative factor (e.g., an order of magnitude), rather than a constant wavenumber or wavelength interval. Furthermore, this definition allows the scales with the largest energy content to be easily identified when a spectrum is plotted over a logarithmic wavenumber or wavelength axis, whereas the standard linear definition of the spectral density can be misleading in this respect. In the appendix, we explain these points in more detail and why in our opinion the logarithmic definition of the spectral density is more appropriate to analyze certain aspects of error growth.

2) DIFFERENCE POTENTIAL ENSTROPHY DIAGNOSTICS

PV on an isentropic level is a tracer even for complex, three-dimensional flows, as long as the flow is conservative. Because of this feature, diabatic and other nonconservative effects can easily be diagnosed in the PV field evolution and a PV-based error metric provides the opportunity to quantify and compare contributions from different physical processes like advection, radiation, convection, etc. to error growth. Following Davies and Didone (2013), such an error metric has been developed in Baumgart et al. (2018, 2019) and here we use a very similar method, but with a few modifications, as laid out in the following paragraphs.

On isentropic levels, the Ertel PV is defined by

\[
\pi = \zeta + f, \quad \frac{\partial}{\partial t} = \frac{\partial}{\partial t} + \mathbf{V} \cdot \nabla \pi - q \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \pi + \nabla \cdot \mathbf{v} + \frac{\mathbf{f}}{\sigma}
\]

(7)

where \( \zeta \) is the vertical component of relative vorticity, \( f \) is the Coriolis parameter, and \( \sigma = -\partial_\rho p/g \) is a measure of static stability (with potential temperature \( \theta \), pressure \( p \), and gravity acceleration \( g \)). The local rate of change of PV can be expressed as

\[
\Delta \pi = -\mathbf{v} \cdot \nabla \pi - q \frac{\partial}{\partial t} + \mathbf{v} \cdot \nabla \pi + \frac{\mathbf{f}}{\sigma}
\]

(8)

where \( \mathbf{v} \) is the horizontal wind and \( \mathbf{V} \) is the horizontal gradient operator, both on isentropic levels, \( q = \theta \) is a diabatic heating rate, and \( r = \zeta \) is the rate of change of vorticity due to momentum forcing.

Diabatic processes in the free atmosphere arise mainly from the release of latent heat due to condensation and freezing and from absorption and emission of longwave radiation, which in a numerical model are computed by parameterization schemes. We will explicitly include the contributions from three parameterization schemes in the diagnostic: the convection scheme (conv), the gridscale precipitation or microphysics scheme (gsp) and the longwave-radiation scheme (lwrad). Baumgart et al. (2019) also considered additional schemes (shortwave radiation, turbulence and gravity wave drag), but only found a very small contribution to error growth. Each scheme computes a heating rate \( q \), which is expressed as a temperature tendency at constant volume in ICON. For this study we decided to neglect the momentum tendency from the convection schemes (i.e., \( r \) is set to zero) because first the Plant–Craig scheme which uses the Kain–Fritsch plume model does not simulate a momentum tendency at all and second for the Tiedtke–Bechtold scheme we found a negligible impact compared to the heating rate contribution.

The advection term in the PV tendency equation, Eq. (8), can be split up into contributions from the rotational and divergent winds by applying the Helmholtz decomposition to the horizontal wind \( \mathbf{v} \), i.e.,

\[
\mathbf{v} = \mathbf{v}_{\text{rot}} + \mathbf{v}_{\text{div}},
\]

(9)

with a divergence-free \( \mathbf{v}_{\text{rot}} \) and a curl-free \( \mathbf{v}_{\text{div}} \). In contrast to Baumgart et al. (2019) we do not further split the rotational wind into contributions from PV anomalies on different vertical levels by means of piecewise PV inversion, since it has been shown that tropospheric-deep interactions do not significantly contribute to error growth on average.

To arrive at a process-based error growth metric, we start by defining the domain-integrated PV error over an area \( A \) as
\[ \Delta P = \frac{1}{2} \int dA (\Delta \pi)^2, \]  
\hspace{1cm} (10) \]

which we will refer to as difference potential enstrophy. \( \Delta \pi = \pi_i - \pi_j \) denotes the PV difference between two ensemble members \( i \) and \( j \) on the same isentropic level. Based on the PV tendency Eq. (8) and the Helmholtz-decomposition (9), the tendency \( T \) of the difference potential enstrophy (10) can be decomposed into different contributions,
\[ \partial_t \Delta P =: T = T^{\text{rot}} + T^{\text{div}} + \sum_{\text{par}} T^{\text{par}} + T^{\text{res}}, \]  
\hspace{1cm} (11) \]

where “rot” and “div” denote contributions from advection by the rotational and divergent horizontal wind, respectively, and \( \text{par} \in \{\text{conv, gsp, lwrad}\} \) denotes the contributions from diabatic heating obtained from the convection, gridscale precipitation, and longwave radiation parameterizations. All other processes contribute to a residual term (res), which is calculated implicitly by computing \( \partial_t \Delta P \) from the data and solving for \( T^{\text{res}} \) in (11). The residual thus contains contributions from other parameterizations, the convective momentum transport (if simulated), contributions from numerical diffusion (Baumgart et al. 2018, section 3c), advection of PV errors through the lateral boundaries of the integration area, and numerical errors due to interpolation and limited spatial and temporal resolution of the output data. By combining (8)–(11), a straightforward calculation yields
\[ T^{\text{rot}} = \int dA \Delta \pi (-\Delta \nabla \cdot \mathbf{v}(\pi)), \]
\[ T^{\text{div}} = \int dA \Delta \pi (-\Delta \mathbf{v}_\text{div} \cdot \mathbf{v}(\pi) + \frac{1}{2} \Delta \pi \mathbf{v} \cdot \langle \mathbf{v}_\text{div} \rangle), \]
\[ T^{\text{par}} = \int dA \Delta \pi (-\Delta q \partial_\theta(\pi) - \langle q \rangle \partial_\theta \Delta \pi + \Delta \pi \partial_\theta(q) + \langle \mathbf{q} \rangle \partial_\theta \Delta \mathbf{q}). \]  
\hspace{1cm} (12) \]

As before, \( \Delta \) denotes the difference between ensemble members \( i \) and \( j \), while \( \langle \cdot \rangle \) denotes the mean of the two members. Note that the terms \( \langle \mathbf{v}(\pi)_\text{div} \rangle \cdot \mathbf{v} \mathbf{\Delta} \pi \) and \( \langle \mathbf{v}(\pi) \rangle \cdot \mathbf{\Delta} \mathbf{v} \mathbf{\pi} \mathbf{\lambda} \) have been transformed by two-dimensional integration by parts, which results in a zero contribution in \( T^{\text{rot}} \) and the noted second summand in \( T^{\text{div}} \). In addition the integration by parts leads to a boundary flux term of \( (\Delta \pi)^2 \) (i.e., a spatial redistribution of error) which is small and in our formulation appears as part of the residual.

Technical details of the tendency calculation (12) are given as follows. Basic atmospheric variables and the accumulated temperature increments from the parameterizations are output every hour on a 1° latitude–longitude grid on model levels (terrain-following height-based levels in ICON). A tendency \( q \) is then calculated via centered time differences of the accumulated temperature increments. Then, \( q, u, v, \) and \( \pi \) are linearly interpolated to isentropic levels. Occasionally at a certain grid point and time step the interpolation on a certain \( \theta \) level is not unique due to localized dry-static instability, which occurs with a frequency of about 0.01% near the tropopause. These points are masked out from the integration area. The height of isentropic levels strongly varies with season and slopes down toward the equator. Since it is our intention to investigate error growth at the tropopause level in the midlatitudes, the choice of the \( \theta \) level to evaluate (12) must be adjusted accordingly. To achieve this, the midlatitude integration area from 40° to 60°N/S is divided into eight 5°-wide stripes. For each case and each stripe the isentropic level (1 K resolution) is chosen such that the absolute value of the reference mean PV on this stripe is closest to 2 PV units, which leads to \( \theta \) levels ranging from 302 to 341 K. The reference PV is computed for each case as the ensemble-average PV of the 0.1% experiment, temporally averaged over the first 7 days (which is the time interval that is analyzed later). The Helmholtz decomposition is computed globally with a spectral method (using windpharm; Dawson 2016) from the horizontal wind that was interpolated onto isentropic levels. For this purpose, gaps due to statically unstable grid points and intersections with the ground are filled using bilinear horizontal interpolation and with the surface value, respectively.

So far, difference potential enstrophy and the decomposition of its tendency (12) are defined between two simulations (two ensemble members of the same experiment and case) and for one spatial area \( A \). To get one time series per experiment and process, we average (10) and (12) over all 12 cases, all 10 combinations of ensemble members, and the 8 latitude stripes on both hemispheres. Applying these averages (denoted with the overbar), a process-specific growth rate \( R \) is computed for each experiment via
\[ R^{\text{proc}} := \frac{T^{\text{proc}}}{\Delta P}, \]  
\hspace{1cm} (13) \]

where proc \( \in \{\text{rot, div, conv, gsp, lwrad, res}\} \).

By construction, the PV-based diagnostics presented here can only quantify the direct impacts from the parameterization schemes at the chosen vertical level (here the tropopause) due to their heating rates, which modify the stratification and hence the PV. This heating can however also cause a secondary circulation in the resolved flow (e.g., an updraft) and in turn modify the divergent and rotational components of the diagnostic, which can no longer be attributed to a certain parameterization scheme. In particular, the contribution from the gridscale precipitation scheme to error growth is probably underestimated, because its direct heating impact is small at the tropopause level, whereas the (vertical) advection of the heated air is explicitly simulated in the resolved flow (while being mostly implicit in the convection scheme). We expect however to detect this secondary circulation mainly in the divergent component of the diagnostic (section 3f in Wirth et al. 2018; Sánchez et al. 2020).

3. Results

3a. Predictability time comparison

We begin by comparing and analyzing the predictability times for the different experiments as defined by (3) for two different thresholds \( \alpha \) (Fig. 1). For the \( \alpha = 0.5 \) threshold the estimate of the intrinsic limit is 12.8 days, for \( \alpha = 0.8 \) it equals 17.1 days. These numbers agree well with our earlier study (Selz 2019) and also with recent results from global convection-permitting experiments (Judt 2020, Fig. 2 therein).
The experiments with larger initial condition perturbations now put these intrinsic limit estimates into context with respect to current practical predictability: The 100%S experiment results in a predictability time of 8.6 days for $\alpha = 0.5$. The other experiments with stochastic convection lie in between and show a fairly linear increase in predictability time with relative initial condition error reduction. This is consistent with a constant, amplitude-independent error growth rate early in the forecast. At the 10% level the predictability time reaches 12.1 days ($\alpha = 0.5$), close to the intrinsic limit (12.8 days), and the curve levels off. Thus, the hundredfold decrease in initial condition uncertainty from 10% to 0.1% only leads to 0.7 days of additional predictability, which is similar to the predictability gain from the only twofold decrease from 20% to 10%. These results not only provide an estimate of the possible gain through initial condition perfection of about 4 forecast days, but also reveal that this gain could be basically achieved by reducing the initial condition errors to about 10% of their current magnitude. Further reduction will produce diminishing returns and very little overall benefit.

When comparing the two thresholds to each other, the predictability time curves appear to be similar but shifted by about 4 days toward longer times for the higher threshold. The most striking difference is a clear increase in case-to-case variability as indicated by the 95% confidence intervals. This reflects the fact that the detection of smaller remnants of predictability requires larger sample sizes. To relate the two DKE predictability thresholds to a more conventional forecast quality metric, we computed the corresponding geopotential anomaly correlations (acc). We find that the 0.8-DKE threshold compares to an acc of 0.26 and the 0.5-DKE threshold represents an acc of 0.60 and thus a more useful forecast. Note that ECMWF defines the predictability time with an acc threshold of 0.8 as one of their headline scores, which is an even higher standard of forecast quality (www.ecmwf.int).

The addition of singular vectors to the initial condition uncertainty in the 100%Ssv experiment decreases the predictability time by about 1 day compared to the corresponding experiment without the singular vectors (100%S). Recall that singular vectors are currently part of the ECMWF ensemble forecasting system and might be interpreted as accounting for an underestimation of the initial condition uncertainty derived from the EDA system. In this case the distance to the intrinsic limit and the possible gain through initial condition perfection increases to about 5 days. However, the singular vectors may also account for model errors and limitations in the model uncertainty representation in which case their shortening of the predictability time is unrelated to the initial conditions.

For reference, we also added the predictability times of the ECMWF ensemble forecasting system (ENS) computed with the same metric and from a similar case sample. The ECMWF forecasts are started from the 100% initial condition uncertainty plus the singular vectors and in addition run the SPPT model uncertainty representation scheme during the entire simulation. This scheme adds additional variability and further reduces the predictability time by about 0.4–0.9 days compared to the 100%Ssv ICON experiment. Recall that Buizza and Leutbecher (2015) found predictive skill for up to

---

1 For this, we correlate the 300 hPa geopotential anomaly of members 2–5 to the first member at the forecast time when the DKE threshold is crossed, which is done for every case and experiment and is then averaged. The anomaly is computed relative to the temporal average from the 31 day simulation interval.

2 It was necessary to use the extended range forecasts, which run for 46 days (instead of only 15 days) but which are only started on Mondays and Thursdays. Hence the initial dates for the 12 cases had to be shifted by 0–2 days and represent a slightly different sample. To make the uncertainty comparable, we also analyzed only 5 random members of the 50-member ECMWF ensemble.
23 days for the ECMWF system based on probabilistic metrics and a much larger sample, while with our metric and thresholds the predictability time is only 7.1 and 10.6 days, respectively, and well below the corresponding estimate for the intrinsic limit.

The three control experiments that use the deterministic convection scheme (i.e., 100%Shr, 10%Shr, 0.1%Shr) show longer predictability than their stochastic counterparts, as expected, and the difference becomes larger for smaller initial condition uncertainty. For \( \alpha = 0.5 \) the difference is highly significant for the 10% and 0.1% experiments \((p < 5 \times 10^{-5})\) and the increase in predictability time (probable overconfidence) of the deterministic convection experiments equals 1 day or about 8%. The difference at 100% initial condition uncertainty equals 0.3 days or 3.8%, which also reaches statistical significance but at a lower level \((p = 0.02)\). This means that the choice of the convection scheme has only a small influence on predictability if current initial condition perturbations are used, but leads to much larger differences when the intrinsic limit is approached and may therefore be more crucial for precise estimates of this limit. For the \( \alpha = 0.8 \) threshold the differences are similar in absolute magnitude but do not reach significance due to larger case-to-case variability \((p > 0.1)\).

The two experiments with the higher 20 km resolution (i.e., 100%Shr, 0.1%Shr) only show a small deviation from the corresponding experiments at the lower 40 km resolution: The predictability time for \( \alpha = 0.5 \) is a bit longer (by 0.2 days) at the higher resolution for the 100% initial condition uncertainty and a bit shorter (by 0.1 days) for the 0.1% initial condition uncertainty. Given the case-to-case variability, neither of these differences reaches statistical significance \((p > 0.25)\). Note that there are no data from the higher-resolution experiments for the larger 0.8 threshold because the simulation length had to be limited due to computational cost.

Interesting differences between the higher and lower-resolution experiments appear for smaller thresholds, i.e., for higher demands on the forecast quality. For \( \alpha = 0.1 \) (which corresponds to a geopotential acc of 0.96), both the 100%Shr and the 0.1%Shr higher-resolution experiments show a predictability time that is shorter by 0.1 and 0.2 days, respectively (not shown), and due the reduced case-to-case variability at this lower threshold these differences reach statistical significance \((p < 0.02)\). This small but significant decrease in predictability time suggests that there is slightly faster error growth initially with higher resolution, but this difference is too small to be detectable at the \( \alpha = 0.5 \) threshold with our sample size. A further correlation analysis (not shown) reveals that the initial error growth acceleration that comes with the resolution increase for the 0.1% experiments is related to the amount of gridscale precipitation at the initial time but not with the amount of convective precipitation. This indicates that the resolution adaptation of the stochastic convection scheme works reasonably well, but better resolved mesoscale details related to stratiform precipitation events such as warm conveyor belts probably lead to faster error growth early in the forecast. For the 100% initial condition uncertainty, the resolution effect is smaller and is not correlated to stratiform nor to convective precipitation.

### b. Process-based analysis

The difference potential enstrophy growth rates, derived in section 2c(2), are now used to gain insight into the physical processes responsible for error growth and predictability limits in the different experiments. The results are shown in Fig. 2, where for each experiment the process-specific error growth rates are plotted over forecast lead time for the first 7 days. The higher-resolution experiments are omitted since the results are similar to the corresponding standard-resolution runs.

For the stochastic convection experiments with large initial condition perturbations (100%S, 50%S), the figure shows that error growth is dominated by the rotational wind, which is consistent with results from previous applications of the diagnostic to operational forecasts (Baumgart et al. 2018; Baumgart and Riemer 2019). The experiment with singular vectors included (100%Ssv) is qualitatively similar but shows an even higher growth rate in the rotational wind component for the first 2 days, which identifies the error growth acceleration generated by the singular vectors. In all of these three experiments, contributions from the divergent wind and from parameterizations are much smaller and of similar magnitude to each other. Among the parameterization schemes, longwave radiation has by far the largest contribution, while direct contributions from the convection scheme are barely visible.

With decreasing initial condition uncertainties, this situation changes, especially in the first 1–2 days: The contribution from the rotational wind decreases, while the contributions from divergent wind and from parameterizations increase. In the 10%S and 0.1%S experiments, the parameterizations dominate error growth for about the first 18 h, followed by a peak in the divergent wind contribution. The rotational wind contribution builds up over the first 4 days and it becomes dominant after about 2 days. The relative contributions from the parameterization schemes also change, with the convection scheme being most important early on, until after about 1 day the longwave radiation scheme dominates again. The (direct) contribution from the gridscale precipitation remains relatively small over the 7 days.

The reference experiments with deterministic convection (100%D, 10%D, 0.1%D) appear qualitatively similar to their stochastic counterparts. For large initial condition uncertainty there is almost no difference at all (100%S versus 100%D), which is consistent with the similar predictability times of the two experiments. This indicates that error growth from current initial condition uncertainties is quite insensitive to the choice of the convection scheme. On the other hand, for the small initial condition error experiments (10%, 0.1%), there are some interesting differences. The experiments with deterministic convection show a weaker initial peak in the parameterization growth rate, followed by a slower decay and thus a slightly larger parameterization growth rate after 1 day compared to the respective stochastic experiments. The peak in the divergent contribution is weaker with deterministic convection, while the rotational contribution is largely unaffected by the choice of scheme. Finally, the relative contribution from the deterministic convection scheme also is smaller initially but remains relevant for a longer time. These results
show that if the initial perturbation amplitude is small, even a deterministic convection scheme will strongly amplify error growth. However, as pointed out by Zhang et al. (2003), this may at least in part be due to an unintended and nonphysical variability resulting from discrete thresholds and switches in the scheme which can quickly amplify very small initial differences. Regardless of that, the error amplification is stronger in the stochastic convection experiments, which eventually results in a reduction of the predictability time estimates compared to deterministic convection (Fig. 1).

The plots also show the residual term (res), which collects contributions from other parameterizations, from diffusion, from advection of error in and out of the domains and from numerical errors in the calculation of the error growth rates. After 1 day, this residual growth rate has values around \(-0.25\) to \(-1.0\) day\(^{-1}\) in all experiments, with some fluctuations. It is negative, which indicates that the processes which are not considered explicitly act collectively as an error sink with numerical and parameterized diffusion likely having the largest contribution. In the first day, however, the residual term differs substantially between the experiments: It is larger by value (more negative) in the 100% experiments, which can be explained by increased diffusion due to inaccuracies and excess small-scale variability in the initial conditions caused by the interpolation from ECMWF EDA to ICON with about half the resolution. In contrast, for the 0.1% experiments, the residual shows a large positive peak initially. A further decomposition of the residual (not shown) revealed that this peak is not caused by the parameterization schemes that have not been considered explicitly, nor by advection effects across the domain boundaries. It is therefore likely caused by an underestimation of the explicitly considered parameterized processes due to spatial and temporal discretization errors in the calculation of the growth rates. For the very fast initial error growth in the 0.1% experiments the output frequency of 1 h and/or the output resolution of 1° is probably too low to compute the process-specific growth rates accurately. Nevertheless, these potential errors do not affect the conclusions discussed here, since they indicate that the differences between large and small initial condition uncertainty experiments are likely even larger during the first day.

In agreement with the predictability times, the potential enstrophy diagnostics confirm that forecasts from the 10% initial condition uncertainty level are already approaching the intrinsic predictability limit, since the convection scheme and divergent wind contributions already dominate the error growth initially, and a further reduction of the initial condition error only leads to a strong increase in the convection-scheme error growth rate, which offsets the gains from the improved initial conditions. Other than this trend, there is very little difference between the 10%S and the 0.1%S experiments.
As pointed out at the end of section 2c(2), the PV-based diagnostics used in our study can only quantify the direct effects of the parameterization schemes on the PV. This means that some fraction of the contributions from the parameterizations’ heating rates appears in the divergent component. Especially the contribution from the gridscale precipitation may be underrepresented by the direct effect, since most of the heating occurs well below the tropopause. It is therefore likely that some part of the increasing divergent component with decreasing initial condition uncertainty is related to the gridscale precipitation scheme, suggesting that more large-scale structures of latent heat release like warm conveyor belts may also contribute to limiting predictability.

The separation of the potential enstrophy error growth rates into different processes does not directly involve any spatial-scale separation (e.g., by means of a Fourier transform). Nevertheless, characteristic spatial scales can be inferred from the typical scale of the underlying processes. Convective heating rates, especially when generated with the stochastic convection scheme, show high variability close to the gridscale. Furthermore, the divergent wind has much reduced variability on synoptic and planetary scales compared to the rotational wind. It can therefore be anticipated that if error growth is mainly triggered by convection and divergent wind, these errors develop on small scales. On the other hand, if error growth is mainly triggered by the rotational wind, the errors likely develop on larger, synoptic scales. Thus the temporal development of the leading error growth mechanism in the 10% and 0.1% experiments from convection to divergent wind to rotational wind also suggests an upscale error from the 10% and 0.1% experiments from convection to divergent wind to rotational wind also suggests an upscale error growth process and an adjustment process to geostrophic balance (e.g., Zhang et al. 2007; Bierdel et al. 2017).

c. Spectral analysis

The novel and complex PV diagnostics will now be complemented with the simpler, “classic” spectral analysis of (difference) kinetic energy. Figure 3 shows the time development and growth rates of the DKE as a function of spatial scale for the first 7 days. Recall that the DKE spectral density (5) has been defined with respect to the logarithm of the wavenumber $k$ according to (6) instead of the standard linear definition to better represent the scale range where the maximum error energy occurs. The reasoning is explained in detail in the appendix, which also includes an alternative version of Fig. 3 that uses the standard linear definition (Fig. A2).

From Fig. 3 it can be seen that after an initial adjustment phase, all experiments enter a stage of nearly constant exponential growth across the synoptic- and planetary-scale range with growth rates of about 0.5–1.0 day$^{-1}$, which can be identified by the constant spacing of consecutive DKE curves. This stage begins early in the forecast for experiments with large initial condition uncertainty and a few days later for experiments with small initial condition uncertainty. It lasts several days until the error growth slows down on successively larger scales due to the saturation limit getting closer (for the small initial perturbation experiments this slowdown is after the 7 day plot range). The constant error growth rate across scales is often referred to as “up-amplitude” error growth (Maps et al. 2008; Durran and Gingrich 2014; Judt 2018), in contrast to the “upscale” growth described by Lorenz (1969). However, Rotunno and Snyder (2008) have shown that the Lorenz model does predict the up-amplitude error growth behavior if it is modified to use the correct large-scale $k^{-3}$ background energy spectrum, rather than the flatter $k^{-5/3}$ spectrum. They further argue that the up-amplitude error growth can be interpreted as a projection onto the fastest growing spectral eigenmode, which is the net result of simultaneous upscale and downscale interactions. Although the DKE growth rate during this phase is initially independent of scale, the decrease of the background energy toward smaller scales means that those scales saturate sooner, leading to the peak in error energy moving upscale over time (see red lines in the figure).

While all experiments eventually enter the up-amplitude growth phase, the initial error growth behavior depends strongly on the amplitude of the initial condition perturbations. Generally speaking, the smaller the initial condition perturbations are, the faster the initial error growth is, and the longer the stage of increased error growth lasts until the stage of steady up-amplitude growth is reached. However, the initial growth rates of the 10% and 0.1% experiments are not only higher compared to the experiments with larger initial perturbations, but clearly peak at much smaller scales. Hence the error maxima move downscale during the first few hours, before moving upscale again, as shown by the red lines in the figure. In simplified turbulence models, rapid error growth on small scales has been linked to a shallower $k^{-5/3}$ spectral slope (Rotunno and Snyder 2008; Durran and Gingrich 2014), but in our experiments it occurs in the context of a $k^{-3}$ slope of the background KE spectrum (see below) and small initial perturbations. The fast initial growth at small scales is likely generated by diabatic processes close to the grid scale, as suggested by the PV diagnostics in the previous section. In contrast, in the experiments with large initial perturbations, the error maxima move upscale right from the beginning. This reinforces the previous conclusion that when the initial perturbation is decreased to about 10%, a transition occurs from a large-scale barotropic error growth regime to a small-scale diabatic growth regime, where the benefits of further improvements to the initial conditions are much smaller and eventually zero. Note that the scale of the error maxima only becomes obvious when the spectral density is defined per unit log $k$ (as in Fig. 3). In the standard linear definition (as in Fig. A2) the DKE spectra show an almost flat tail toward the large scales with no distinct maximum and the identification of the transition is much more subtle since it appears only in the elongation of the flat tail toward smaller scales.

In all experiments, the average background kinetic energy (thick black line in Fig. 3) appears very similar and follows a $k^{-3}$ power law from about 4000 km downscale. At the smallest scales plotted, numerical diffusion steepens the slope a bit.  

\[3 \text{Technically it is a } k^{-2} \text{ power law for the log-defined KE (see appendix).}\]
A transition to a flatter $k^{-5/3}$ spectrum, which should occur at around 400 km (Nastrom and Gage 1985), cannot be seen since it would require simulations with a much higher resolution than the one we used. Although not evident from the figure, the KE spectra of experiments with stochastic convection have slightly more energy at scales below 1000 km (up to 25% near the grid scale) compared to the deterministic convection experiments. This is consistent with the design of the stochastic convection scheme, which introduces more variability close to the grid scale, which eventually energizes the upper tropospheric wind fields, enhancing both the background and error kinetic energies. The resulting slightly shallower slope of the KE spectrum at small scales could be responsible for the small remaining sensitivity of the predictability limit to the choice of the convection scheme at 100% initial uncertainty (Fig. 1), although the potential enstrophy diagnostics (Fig. 2) did not show any direct contribution of the convection scheme to error growth.

It is interesting to contrast the influence of small-scale diabatic processes with that of the singular vector perturbations. Figure 3 shows that the accelerated growth of the 100%Ssv experiment (compared to 100%S) occurs on synoptic and planetary scales, and does not produce a downscale shift of the error maximum initially. The singular vector induced growth slows down after
about 2 days, as expected from their optimization interval at ECMWF. In the PV diagnostics we saw an analogous increase in the rotational component, which further confirms that error growth on larger scales (synoptic to planetary) is driven mainly by the rotational wind.

Finally, we note that in the experiments with large initial condition uncertainty, there is a slight decay of difference kinetic energy (negative growth rates) at small scales and a lower growth rate in general during the first 12 h. This anomaly is caused by the generation of the initial conditions through interpolation from the ECMWF system to the ICON model, which includes a decrease in resolution from 18 to 40 km. The interpolation leads to an excess kinetic energy in the initial conditions at small scales compared to the model spectrum of the lower ICON resolution, which is removed by diffusion during the first 12 h of simulation time. The higher-resolution reference experiments have a similar grid size compared to the ECMWF system and slightly more energy at the small scales compared to the standard resolution experiments (not shown). Consequently, the 100%Shr experiment does not experience initial error decay and the errors grow right from the beginning. Other than that the higher-resolution experiments are largely similar to the corresponding experiments at the standard resolution (not shown).

4. Summary and discussion
a. Summary

In this study we investigated the transition of midlatitude predictability from the current practical limit to the intrinsic limit, as the accuracy of the initial conditions is improved. For this purpose we performed global ensemble simulations with a numerical weather prediction model (ICON) at 40 km resolution started with rescaled initial condition uncertainty amplitudes, ranging from current estimates to uncertainties that are 1000 times smaller. The model includes a stochastic convection scheme (Plant and Craig 2008) to compensate for the underestimation of convective variability and subsequent upscale error growth that can occur when convection is unresolved and deterministically parameterized. The much lower cost compared to using convection-permitting resolution made it possible to conduct experiments with 5 different initial condition uncertainty rescale factors based on 12 cases and 5-member ensembles and to conduct additionally reference experiments with deterministic convection and singular vectors. To provide a basic check on the resolution sensitivity, two of the experiments were repeated at twice the resolution (i.e., 20 km) but with only about half the forecast lead time. In addition to estimating predictability times, we used novel process-specific PV diagnostics and standard kinetic energy spectral analysis to identify key characteristics of early-stage error growth as a function of the amplitude of the initial condition uncertainty.

We obtain an estimate for the intrinsic limit of about 2 weeks, consistent with many earlier studies, but of course the precise value will depend on variable, level, metric, and the requirements of a potential user. For 300-hPa-difference kinetic energy we found that for the intrinsic limit experiment, the DKE reaches half the climatological variance after 12.8 days, which corresponds to a geopotential anomaly correlation of approximately 0.6. This suggests that the predictability limit of current forecasts could potentially be extended by 4–5 days with perfecting the initial conditions, depending on whether the effects of the singular vectors are included or not. The sensitivity of these estimates to a doubling of the resolution from 40 to 20 km was small and insignificant. A 90% reduction of the current initial condition uncertainty is sufficient to bring the predictability close to the intrinsic limit and further reductions will show little additional benefit with respect to forecasts of upper-tropospheric flow patterns. Given that the rate of improvement of operational forecasts has been about 1 forecast day per decade (Bauer et al. 2015; Stern and Davidson 2015) for the last 50 years, this implies that the intrinsic limit could be reached in about 40–50 years if this rate were to continue.

A decomposition of potential enstrophy error growth rates into specific processes showed very pronounced differences in the initial stage error growth for different amplitudes of the initial condition uncertainty: Error growth from current initial condition uncertainty estimates is dominated by uncertainty in advection by the rotational wind. In contrast, if the initial condition uncertainty is very small (0.1% of its current value), contributions from the convection scheme become dominant initially, followed by a dominance of uncertainty in advection due to the divergent wind. The initially small rotational component builds up over time and dominates again after about 2 days. This is consistent with our earlier study (Baumgart et al. 2019) and the three-stage model of upscale error growth and geostrophic adjustment (Zhang et al. 2007). The additional experiments with initial condition uncertainties between these two extremes reveal where the transition occurs. The 50% experiment shows dominance of the rotational component, similar to the unchanged initial condition experiment, while the 10% experiment shows a three-stage upscale growth, similar to the 0.1% result. The 20% experiment lies in between, with similar contributions from divergent wind, rotational wind, and parameterizations initially.

A spectral analysis of error growth using difference kinetic energy also demonstrates distinct features depending on the initial condition uncertainty, consistent with our process-based analysis. Experiments with small initial perturbations show a fast downscale transition of the error maximum during the first few hours, which is followed by upscale propagation of the peak error energy from there. This resembles error growth characteristics found in simplified 2D-turbulence models with a $k^{-5/3}$ regime although the spectral slope in our simulations stays close to $k^{-3}$ down to the effective resolution of $O(100)$ km. After a few days, a stage of quasi-exponential growth on synoptic and planetary scales is entered. Experiments with large initial condition uncertainty show this quasi-exponential growth behavior almost from the start.

The reference experiments using the deterministic Tiedtke–Bechtold convection scheme simulate a 5%–10% or 1 day longer intrinsic predictability time relative to experiments with stochastic convection. The difference becomes more pronounced if the required forecast accuracy is increased (i.e., lower saturation threshold). The predictability difference between deterministic
and stochastic convection experiments, and hence the sensitivity to the choice of the convection scheme, increases with decreasing initial perturbation amplitude confirming that moist processes at small scales become more important in limiting unpredictability when the initial condition uncertainty is smaller. In a qualitative sense, we found little difference between the stochastic and deterministic experiments. Quantitatively, the initial error growth contribution from convection is somewhat smaller but longer lasting in the deterministic experiments, which indicates a delayed upscale error growth process in the experiments with small initial condition uncertainties.

b. Assumptions and limitations

Our study aims to estimate the predictability of the real atmosphere, including all its scales. It is however based on a perfect model assumption, but uses a model that is imperfect. With the adoption of the stochastic convection scheme we have attempted to account for the most relevant contributions from processes and scales that are not (well) resolved at a resolution of 40 km. The quantitative predictability estimates we obtain should be reasonably accurate if the following assumptions are valid.

1) The stochastic convection scheme adapts to the model resolution and projects the correct convective variability onto the resolved scales.
2) Error growth in regions with stratiform precipitation and in dry regions is either already reasonably well simulated at 40 km resolution or has a slower growth rate compared to regions with convective activity and hence does not significantly contribute to the intrinsic limit.
3) Error growth on synoptic to planetary scales is reasonably well simulated at 40 km resolution.

We discuss some details about these assumptions and their validity below.

The scale adaptivity of the Plant–Craig convection scheme is a direct consequence of the underlying theory of convective variability and has been confirmed in numerical simulations (Keane et al. 2014). We have investigated the scale dependence of error growth in a previous study (Selz and Craig 2015a), and found that error growth is well represented by the stochastic convection scheme for a range of resolutions compared to a 2.8-km-reference experiment with explicit convection. We note, however, that this study considered only a single case with summertime convection over continental Europe. Furthermore, Weyn and Durran (2018) showed that even at grid sizes of 1–2 km error growth from convection could still be sensitive to resolution, so that even “convection-permitting” simulations with grid sizes of O(2–4)km still may suffer from a predictability overconfidence.

Another major concern with respect to the resolution used in our study is the missing transition to a flatter $k^{-5/3}$ spectrum, which should occur roughly at 400 km wavelength. In simplified models following and extending the 2D-turbulence theory of Lorenz (1969) (e.g., Rotunno and Snyder 2008; Durran and Gingerich 2014; Leung et al. 2020) this transition plays an essential role in intrinsically limiting predictability, since in the $k^{-5/3}$ range an almost immediate error saturation occurs that quickly propagates upscale and into the $k^{-3}$ range. On the other hand, a transition to a $k^{-5/3}$ spectrum does not seem to be that crucial in more realistic simulations conducted with complex weather prediction models. For example, the experiments of Judt (2018, 2020) at 4 km resolution did represent the $k^{-5/3}$ range well but did not find predictability times that contradict earlier studies and agree well with the results of this study. Furthermore, Zhang et al. (2019) reported little sensitivity with respect to a resolution increase from 9 to 3 km for their winter case, although this increase has been accompanied by a significant flattening of the mesoscale spectral slope. Finally, Lloveras et al. (2022) recently reported that error growth from small-amplitude perturbations is insensitive to the slope of the background kinetic energy spectrum in convection-permitting simulations of idealized midlatitude cyclones.

Diabatic processes, especially latent heat release, may explain this discrepancy since they are ignored in simplified turbulence models, while identified as essential for the error growth that occurs in realistic simulations, regardless of whether convection is parameterized or not (e.g., Zhang et al. 2003; Hohenegger and Schär 2007; Selz and Craig 2015b). The simulation of latent heat release in weather prediction models likely triggers fast error growth on scales near the effectively model resolution, which is largely insensitive to the slope of the kinetic energy spectrum at those scales. Indeed the 0.1% experiments from our study also feature fast initial growth rates at the smallest resolved scales, which would require the presence of a $k^{-5/3}$ mesoscale range in the idealized turbulence models. Finally, there is evidence that the flattening of the slope at the mesoscales may itself be a consequence of moist processes (Sun and Zhang 2016; Selz et al. 2019), and that the transition to a flatter spectrum occurs at larger scales and is more pronounced in areas with precipitation and latent heat release. In this way the idealized turbulence models may have incorporated the effects of moist processes on error growth indirectly through the value of the slope of the kinetic energy spectrum. A somewhat different perspective was given by Lloveras et al. (2022), who argued that the localized nature of error growth in moist convection violates the homogeneity assumption of the turbulence models and leads to broad errors in spectral space which makes them insensitive to the presence of a $k^{-5/3}$ range.

In the stages of error growth where error amplitudes peak on synoptic scales, the growth is basically decoupled from moist processes. However, even in this regime error growth could be highly sensitive to model resolution, as demonstrated by Morss et al. (2009), who found a 1–2 day shortening of the predictability time with every doubling of resolution, using a quasigeostrophic model. We on the other hand found relatively little sensitivity with respect to doubling the resolution from 40 to 20 km: Only for low saturation thresholds (i.e., high demands on the forecast accuracy) was the change in predictability statistically significant, and the difference of 0.1–0.2 days is an order of magnitude smaller than the sensitivity found by Morss et al. (2009). In the 0.1% experiments this sensitivity seems to be correlated with the amount of
gridscale precipitation at the initial time. No such correlations could be found for convective precipitation or in the 100% experiment. Although not investigated in detail for this paper, these findings suggest that the most critical of the assumptions listed above is the accurate simulation of moist processes that are not represented by the stochastic convection scheme, such as mesoscale structures in more stratiform ascent regions like warm conveyor belts.

Independent of the amplitude of the initial condition uncertainty, all experiments enter a stage of up-amplitude error growth associated with errors in the advection by the rotational wind and on synoptic to planetary scales. This is consistent with the quasi-2D turbulence argument of Lorenz (1969), and there is little evidence that these processes are not well simulated in current numerical weather prediction models, such as the one used here. The existence of the intrinsic predictability limit, on the other hand, depends on rapid growth of small-scale errors associated with diabatic processes. There is considerably more doubt about the accuracy of current models in simulating these processes, certainly with parameterized convection, and even for kilometer-scale models that only partially resolve convective clouds. However, based on the comparison between the stochastic and deterministic convection experiments we expect the uncertainty to be on the order of 10% with respect to the 2-week intrinsic limit, which reflects the predictability of larger-scale upper-tropospheric flow patterns. Definite answers can probably only be achieved by larger samples of global convection-permitting simulations or even global large-eddy simulations, until the results have converged and a sensitivity to further resolution increases can no longer be found.

Predictability of the midlatitude troposphere can also be influenced by processes from outside. Many related parts of the Earth system evolve on somewhat longer time scales and could provide a low level of predictability beyond 2 weeks even in current forecasting systems (e.g., Buizza and Leutbecher 2015). On subseasonal to seasonal time scales, the most relevant phenomena that could enhance the predictability of the midlatitude troposphere are the stratospheric polar vortex, the tropical Madden-Julian oscillation (MJO), and anomalies in the ocean, sea ice, snow cover and soil moisture [see, e.g., Robertson and Vitart (2018) for an overview]. The ICON model used in our study does simulate most of these systems (see section 2a) and although no ocean model is included, sea surface temperature anomalies and the phase of the MJO are likely reasonably well represented in the initial conditions derived from ECMWF (Vitart and Montani 2010). In this sense interactions from these systems and their influence on midlatitude predictability are included in our simulations, admittedly with varying levels of accuracy. However, these influences are relatively weak, and most likely excluded by our choices of metrics and thresholds, which define a relatively high level of remaining predictability as the cutoff. The investigation of predictability levels closer to the climatological distribution would require a much larger sample of cases and members. It seems likely that this would increase the importance of the interaction processes mentioned above. They may eventually become the dominant contribution to this low-level predictability, which makes their accurate representation crucial. The intrinsic limit estimated from such a study [e.g., along the lines of Buizza and Leutbecher (2015), the “horizon of the intrinsic limit”] will certainly turn out to be much longer than 2 weeks.

This study, as well as related previous work (Selz 2019; Zhang et al. 2019), relates a potential increase in predictability to improvements in the initial conditions only. However, current practical predictability is not only limited by the initial condition uncertainty but is also significantly limited by uncertainties and errors in the model. If these were eliminated the predictive skill would be expected to increase. In the ECMWF system, model uncertainty is represented by the SPPT stochastic parameterization and possibly to some extent by the singular vectors. Both methods combined lead to a reduction of the predictability time of about 2 days compared to the experiments that include the EDA-initial condition uncertainty only (Fig. 1). This suggests that by eliminating model errors the predictability time could be extended by up to 2 days, independent of initial condition improvements, and this improvement potential is not included in the design of our study. The exact number is however largely unclear, since it is unknown to what extent the singular vectors compensate for uncertainties in the model and to what extent errors would grow faster in a perfect model. Finally, we note that a reduction of the initial condition uncertainty in current forecasting systems may be less beneficial than perfect model estimates suggest if the model could not be improved accordingly (Zagar and Szunyogh 2020).

c. Significance of our results for current forecasting systems

A major result of this paper is the estimate that a 90% reduction of the initial condition uncertainty would be sufficient on average to reach the intrinsic limit of predictability. The fact that current initial condition errors immediately trigger error growth on synoptic and planetary scales and in the rotational component of the flow (Figs. 2 and 3) indicates that error reduction is mainly required on those scales to increase current practical predictability, which is in agreement with Durran and Gingrich (2014). These authors further point out that intrinsic predictability is not necessarily limited by “butterflies” (i.e., initial errors on small scales), but by any (small amplitude) error on any scale (also see Sun and Zhang 2016). However, this does not mean that processes on small scales are unimportant. In fact, the results presented here clearly show that especially convection significantly contributes to limiting intrinsic predictability and that the transition where they become the limiting factor occurs at initial condition uncertainties below 10%–20% of current estimates.

An additional finding is that, except for the first day in the experiments with small initial condition uncertainty, the PV diagnostic shows a clear dominance of the longwave radiation scheme over the other parameterized processes with respect to error growth (Fig. 2). The sharp humidity gradients at the tropopause lead to a peak of the cooling rates from outgoing longwave radiation, which in turn increase the PV aloft and
decrease it below (Chagnon et al. 2013). These tendencies then further amplify existing differences in the PV structures at the tropopause and explain the steady positive error growth contribution from the longwave radiation scheme. This effect may be intensified by the presence of high-level clouds due to even stronger radiative cooling. Also additional research on this point is required, our results suggest that even with current initial condition perturbations the accuracy of the moisture distribution at the tropopause and the accurate representation of related radiative processes deserve some emphasis.

The small impact of the stochastic convective tendencies to error growth at present-day levels of the initial condition uncertainty indicates that neither spread nor skill of modern ensemble forecasting systems would benefit from the inclusion of a stochastic convection scheme like Plant and Craig (2008) to the uncertainty representation of the model. Although this is likely true on average and with respect to upper-tropospheric large-scale flow patterns, there is still potential benefit from adopting the stochastic convection scheme for precipitation forecasts, especially with respect to local extremes. Wang et al. (2016), for example, showed that the Plant–Craig scheme greatly improved the frequency of strong precipitation events and at the same time reduced the “too-much–drizzle” problem that deterministic convection schemes often have. In principle, it should be possible to combine a scheme like Plant and Craig (2008) with current methods of model uncertainty representation like SPPT or the recently developed stochastically perturbed parameterizations scheme (SPP; Lang et al. 2021), since they represent different aspects of parameterization-related errors: The Plant–Craig scheme introduces missing variability close to the grid scale that arises from violating the scale separation assumption when model grid boxes are too small to contain a large number of clouds. As a consequence of this violation, fluctuations of convective mass flux around the mean become significant. A similar idea has recently been applied to large eddies in boundary layer turbulence for the use in kilometer-scale models (Hirt et al. 2019). In contrast, schemes like SPPT and SPP address all kinds of uncertainties and errors related to design and simplifications of the parameterization schemes and their variability is introduced on much larger spatial and temporal scales.

This paper investigated predictability on average by considering 12 cases distributed over one entire year and the midlatitudes of both hemispheres. We did not investigate the seasonal, regional, or flow-dependent variability of predictability, which would require a larger sample of simulations. Furthermore, for some specific flow patterns, the conclusions from our study might not apply and we cannot exclude the possibility that influences from the convective scale even today could intrinsically limit predictability locally and lead to extraordinary bad forecasts. A potential mechanism for this was described by Rodwell et al. (2013), who pointed out that bad forecasts over Europe can be associated with strong convective activity over North America 5 days earlier. The results and diagnostics presented here may be helpful for future investigations to answer the vital question of whether significant improvements in forecast skill are also possible for such bust cases, or if their skill is already restricted by a local and transient decrease of the intrinsic predictability limit.

Acknowledgments. The research leading to this paper was carried out as part of the Collaborative Research Center SFB/TRR 165 “Waves to Weather” within project A1: “Upscale impact of diabatic processes from convective to near-hemispheric scale” funded by the German Research Foundation (DFG). The use of ECMWF’s computing and archive facilities is gratefully acknowledged. We thank Chris Snyder and two anonymous reviewers for their careful evaluation of the original manuscript and their insightful comments.

Data availability statement. The ICON output data that are required to reproduce the diagnostics presented in this paper (4.4 TB in total) are available for download. You will find a download script and additional information at https://github.com/wavestoweather/data_download_selz_etal_2022. The ICON model code is restricted software and cannot be made publicly available by us. Interested readers can apply for access at https://code.mpimet.mpg.de/projects/iconpublic. The initial condition data are retrieved from ECMWF’s operational archive, which is also restricted. To gain access, visit www.ecmwf.int.

APPENDIX

Linear Versus Logarithmic Definition of the Spectral Density

The atmosphere and its phenomena are frequently divided into spatial-scale ranges for analysis or simplification. Although many different definitions exist, usually scales are separated by a constant factor rather than a constant wavenumber or wavelength interval, as apparent, for example, in this common definition of scale: planetary (~106 km), synoptic (~103 km), mesos (~102 km), and convective (~101 km). This exponential characteristic of the term "scale" is also reflected by the fact that spectral graphs usually employ a logarithmic x axis, in which the different scale ranges appear equidistant.

Although spectral transforms on limited domains (like the real atmosphere) result in a discrete, countably number of coefficients, spectral energy (or variance) distributions are usually thought to be continuous and expressed as a spectral density \( \hat{e}(k) \), which satisfies the following equation (e.g., Durran et al. 2017):

\[
\mathcal{E} = \int dk \hat{e}(k),
\]

where \( \mathcal{E} \) is the total spatially averaged energy (or sometimes the total spatially integrated energy). Although this definition is widely used and there is nothing wrong with it per se, the resulting spectral energy distribution can be misleading when plotted on logarithmic axes and given the exponential characteristic of our definition of scale. This is
illustrated in Fig. A1a for a difference kinetic energy spectrum from one of our experiments. It shows the typical shape of 1D error spectra, which appears flat at large scales and eventually decreases toward small scales. In the figure, the available scale range is separated into five bins with the width being a constant factor. The energy contained in each bin is then computed using (A1). Interestingly, the majority of the energy is not found at large scales, which the shape of the spectral density suggests, but at scales where the spectrum has already started to decrease. The reason for this counterintuitive result is that on a linear wavenumber axis, the bin widths differ greatly, with the small-scale bins being much larger than the large-scale bins.

This discrepancy between spectral density amplitude and energy content per scale can be avoided by using a redefinition of the spectral density that takes the logarithmic scale axis into account:

\[ E = \int d\log k \tilde{e}(k), \]  
where \( \tilde{e}(k) \) is the logarithmically defined spectral density, which is related to the linear one by
\[ \tilde{e}(k) = k \hat{e}(k). \] (A3)

In Fig. A1b, the result of this recalibration is shown for the difference spectrum example. Instead of being approximately flat at large wavelengths and decaying for small wavelengths, the spectrum now shows a clear maximum and this maximum coincides with the bin that contains the most energy. It thus better represents where the maximum energy is in the context of an exponential scale definition and when plotted on logarithmic scale axis. Although this logarithmic definition (A2) of the spectral density has been suggested previously, most notably by Lorenz (1969) in his classic study, it has not been widely used in the error-growth literature. Also note that 2D spectra of DKE (Judt 2018; Lloveras et al. 2022) feature a clear peak even without the logarithmic definition of the spectral density and the difference between the two definitions is less prominent in that case.

For (absolute) kinetic energy spectra with their \( k^{-3} \) and \( k^{-5/3} \) power-law ranges, the logarithmic redefinition of the spectral energy will reduce the magnitude of these slopes by one, e.g., the \( k^{-3} \) slope would still appear linear, but reduced to a \( k^{-2} \) slope. This does not lead to any qualitative changes, but the decay of energy toward smaller scales appears less pronounced. Although we use the logarithmic definition (A2) in our paper, we continue to refer to the power-law ranges with the original slope, i.e., \( k^{-3} \) and \( k^{-5/3} \), because these are well established terms and changing them would probably cause confusion. The reader should be aware, however, that the slopes of the background kinetic energy spectra in Fig. 3 are actually \( k^{-2} \) and not \( k^{-3} \).

Finally, for better comparison to earlier studies, we provide a plot of the DKE spectra that uses the standard linear definition of the spectral density (Fig. A2). Although both versions of the plot contain the same information, the changes related to the predictability regime transition are visible more clearly in the log version (Fig. 3), as indicated, for example, by the red lines that connect the maxima.

REFERENCES


