Workshop On Diagnostics For Global Weather Prediction, 9 - 12 September 2024

Mark Rodwell, Matthieu Chevallier, Alison Cobb, Rebecca Emerton, Richard Forbes, Estibaliz Gascon, David Lavers, Sarah-Jane Lock, Linus Magnusson, Michael Maier-Gerber, Sebastian Milinski, Inna Polichtchouk, Luise Schulte, Frederic Vitart.

The task of ensemble forecast system development is to reduce uncertainty, while improving statistical reliability. The aim of this workshop was to discuss how practical diagnostics can help with the many facets of this task. The focus was largely on the prediction of day-to-day weather at lead-times 1 – 45 days, rather than temporal-mean anomalies at longer lead-times. The workshop was organised around some questions of relevance for ECMWF strategy. This report is structured around these questions. Section 1 discusses the topic of diagnosing predictability, which largely relates to the "uncertainty" part of the forecaster's task. In sections 2 and 3, the diagnosis of key processes and process errors is discussed. Lately, there has been a rapid development of the use of Artificial Intelligence in forecasting and section 4 investigates why such AI models are so competitive and some diagnostics which can be applied to them. These models are sometimes referred to as machine-learning (ML) models, or data-driven models to distinguish them from traditional physics-driven models. A key thrust of recent research has been the development of much higher resolution physics-driven models, and of the computing infrastructure needed to run them. In section 5, we seek to diagnose the benefits of higher resolution in the context of global numerical weather prediction (NWP). Section 6 discusses the benefits of community diagnostics tools and common datasets.

This report sites examples from the many excellent diagnostic studies presented; for more details, the reader is referred to the workshop portal (https://events.ecmwf.int/event/383/) where the presentation slides, presentation recordings, and posters can be found. In addition to reporting on the workshop, this document aims to record expert opinion from all participants, with a view to informing ECMWF strategy.

1. What is ultimate predictability, and what is holding NWP back?

This session investigated the limits of predictability for day-to-day weather, if we knew the initial conditions almost exactly. It also opened the discussion of what might be needed to achieve this predictability in practice. "Perfect-model" studies for the midlatitudes (Tobias Selz) suggest that the initialisation of large-scale rotational flow is key for reducing the uncertainty of current forecasts. Ultimately, this can perhaps increase the lead-time to a particular level of uncertainty by 3 days. A further day appears possible if the smaller convective scales can be perfectly initialised (Fig. 1). To put the total potential "4 day" result into context, ECMWF's historic target has been to improve forecasts by 1 day per decade. Later, the workshop explored the sensitivity of the result to model formulation and resolution.

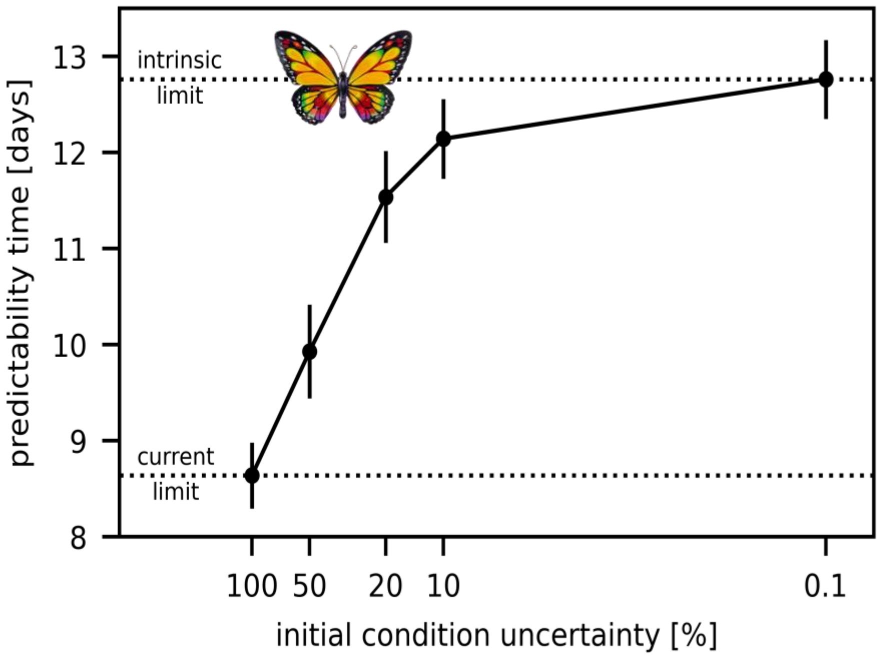


Fig. 1 *Time for ICON ensemble variance in 300hPa kinetic energy to reach 50% of its climatological value, plotted as a function of initial uncertainty scaling. Courtesy Tobias Selz.*

Potential Vorticity (PV) is a useful scalar quantity for studying atmospheric dynamics. This is because dynamical and physical aspects can be neatly separated, tendencies can by partitioned by process, and piecewise inversion can reveal which aspects of the flow are consistent with which PV features. Michael Riemer showed results which suggest that baroclinic development is, on average, subordinate to other error-growth mechanisms (see also Edward Groot). Looking for situations which first encounter the intrinsic limit of predictability could help focus modelling efforts. Fig. 2 shows preliminary results for the mid-latitudes, which identify regions ahead of upper-level troughs that can be prone to convection, but in the presence of substantial Convective Inhibition (CIN). The suggestion, maybe, is that the uncertainty in whether convection does break through this inhibition creates almost a bifurcation in the forecast.

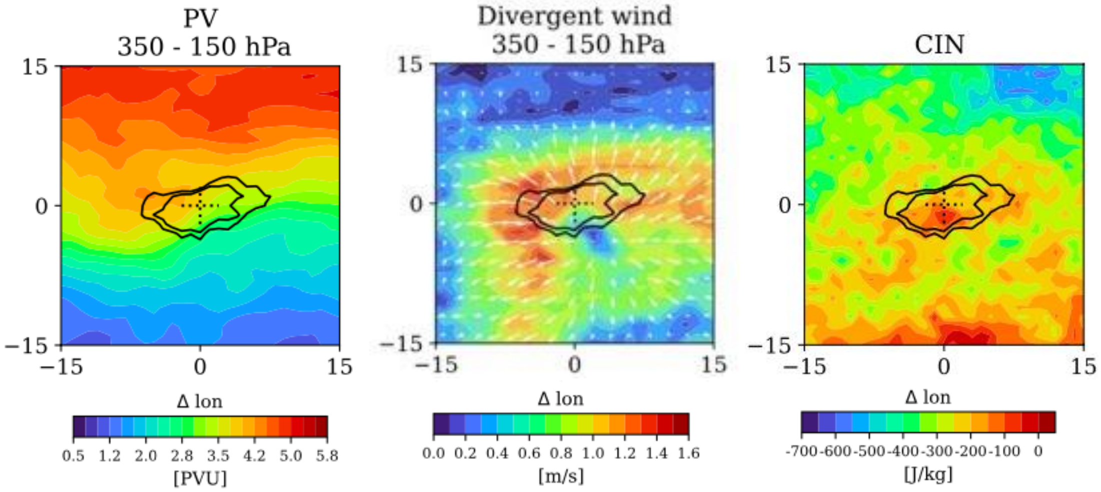


Fig. 2 *Preliminary results showing environmental conditions associated with strongest upscale uncertainty growth (black contours): (a) Upper-level PV, (b) Upper-level divergence, (c) Convective Inhibition (CIN). Courtesy Michael Riemer.*

Before we reach the intrinsic limit of predictability, forecast uncertainty can be reduced through better initialisation. The benefits of improving the observational information available to the data assimilation are graphically demonstrated by the decreasing 40-year trend in European "busts" at day-6 in forecasts made with the (fixed) ERA5 model cycle (Seraphine Hausser; Fig. 3). It is possible that the strength of this trend is partly explained by the fact that such poor forecasts are at the tail of the score distribution, where the relative impact of reduced distributional biases are magnified.

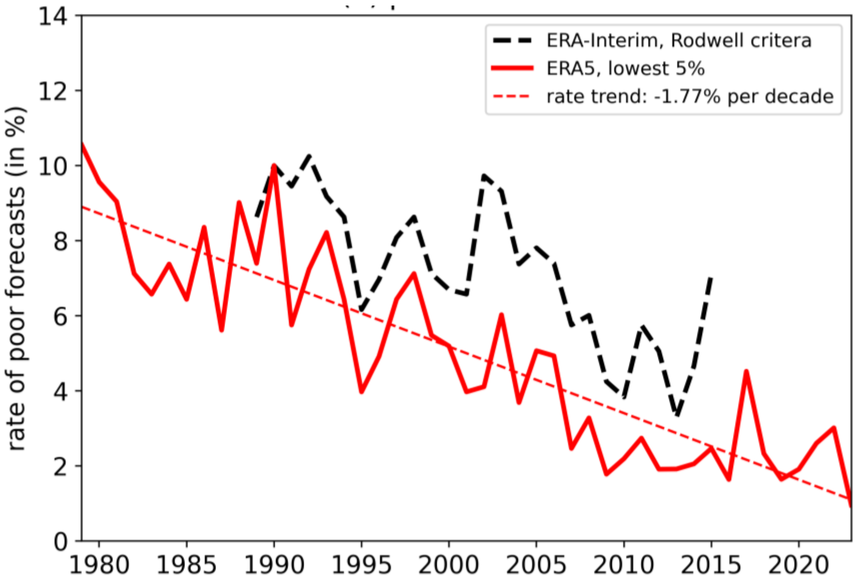


Fig. 3 *Decrease in the number of day-6 European "forecast busts" with time in ERA5 forecasts. There is corresponding rise in the number of very good forecasts. Courtesy Seraphine Hauser.*

In addition to understanding and reducing uncertainty is the need for forecasts to be statistically reliable (Mark Rodwell), and this requires better models – in particular better representation of physical processes and model uncertainty. The identification in forecasts of "barriers" to predictive skill is a good way to identify flow features (in the forecast "storyline"; Christian Grams) which act as intrinsic amplifiers of uncertainty, or which are particularly sensitive to process errors (maybe often both go hand-in-hand). One such flow-feature, or regime, is the warm conveyor belt (WCB; moist airstreams that ascend from the warm sector of an extratropical cyclone into the upper troposphere). Biases and uncertainties in the representation of diabatic heating in WCBs are thought to be important for deficiencies in the subsequent development of larger-scale blocking events (Christian Grams). Mesoscale convective systems (MCSs; regions of thunderstorm activity, typically 100-500km in scale) represent another flow feature associated with the amplification and scale-enlargement of uncertainty. Together, these barriers point to the pivotal role played by diabatic heating, and its model representation, in forecast performance.

For the large scales in the tropics, intrinsic predictability is longer (Hyemi Kim), although with diabatic heating again playing a key role (David Strauss). To help improve predictive skill of weekly-means and seasonal-means, diagnostics need to be additionally applied to other aspects of the Earth System. Although badly needed, predictability studies at these longer ranges are thought to be less meaningful at present, owing to the poor representation of key predictive pathways, such as from the Madden-Julian Oscillation (MJO; the eastward progression of large regions of both enhanced and suppressed tropical rainfall, representing a major component of tropical variability and predictability) and its global teleconnections.

This session demonstrated diagnostic techniques used to quantify predictability and to identify key meteorological features associated with predictability limits. Perhaps the main implication for improving weather forecasts is the need to reduce initialisation uncertainty at large-scales and to improve the representation of diabatic processes.

2. How do we better diagnose model errors and uncertainties?

The quantification of intrinsic predictability assumes a perfect model. In this session, the aim was to explore diagnostics that highlight the imperfections in our models; particularly those associated with convection. Composites from tracking diagnostics (robust across differing tracking packages) show that MCSs account for over half the total precipitation in the tropics and in some extratropical regions such as North America. However, even at very high km-scale resolution models generally produce too intense and localised precipitation (underestimated overall), which is too sensitive to the atmospheric moisture content (Zhe Feng, Fig. 4). The corresponding errors in the magnitude and structure (particularly height) of diabatic heating will have consequences for interaction with the global circulation. MCSs can also be difficult to capture in analyses (Dave Parsons), and this is partly why they can lead to forecast busts. It was noted that the quantification of the MCS's sensitivity to its larger-scale environment, and uncertainty growth associated with its own internal chaotic dynamics (Ziwei Wan) could follow similar diagnostic approaches already used for tropical cyclones.

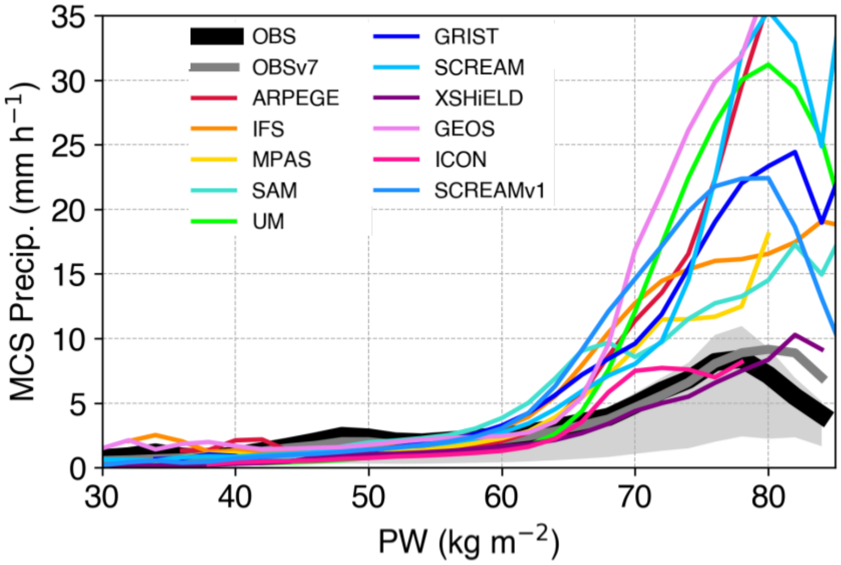


Fig. 4 *Comparison of the sensitivity of MCS precipitation to total precipitable water in several models run at convection-permitting resolution, as part of the DYAMOND initiative. For example,*

the IFS is run at ~4km resolution, with the deep convection parametrization turned off. Observational data are IMERGv6 (black, with shading indicating interquartile range) and IMERGv7 (grey). Except for the IFS, all models are non-hydrostatic. Courtesy Zhe Feng. DOI: 10.22541/essoar.172405876.67413040/v1.

Tropical lower tropospheric circulation biases are prevalent at all time ranges in ECMWF's global Integrated Forecasting System (IFS). These are at least partly attributed to the parametrizations of momentum flux in the boundary layer and shallow convection. As model resolution increases, it is important to quantify the fraction of momentum flux which still requires parametrization (Alessandro Savazzi). Large eddy simulations suggest that this fraction is also sensitive to the degree of organisation (or spatial scale) of the cloud field; something not generally considered in parametrizations. As shown in Fig. 5, at a model resolution of 25km, perhaps 60% (100%) of the flux requires parametrization in situations of strong (weak) organisation.

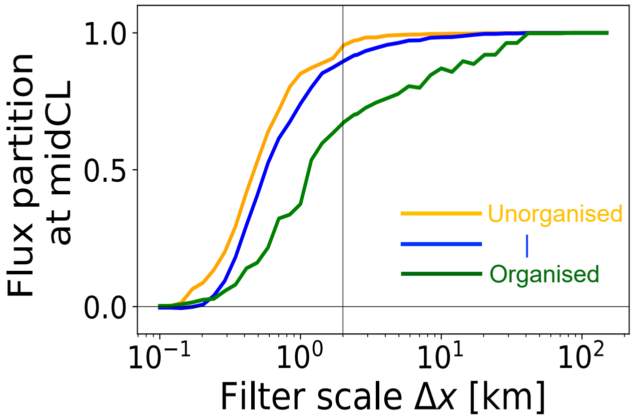


Fig. 5 *Fraction of momentum flux that occurs at shorter scales than a given (model) resolution, for different degrees of cloud field organisation. From DALES large eddy simulations. Courtesy Alessandro Savazzi.*

Diagnosis of forecast errors requires a knowledge of the observed truth. We could make better use of the observations we have, either directly, or through diagnosis of data assimilation, or through nudging approaches (Chihiro Matsukawa). A stumbling block, which might get worse in future, concerns the ownership of observational data, and thus our ability to exchange this, and its associated assimilation feedback information.

There is also a need from Diagnostics (and more generally from Forecasting itself) for new observational information. Here, we list a few examples discussed.

- Aeolus observations provide essential wind information in the tropics, which is complimentary to temperature information (e.g. from Radio Occultation observations), and can help initialise tropical waves (Robin Pilch Kedzierski). Continuation and better spatial coverage of such observations seems desirable.
- More rainfall observations would also be valuable. While we have IMERG (Integrated Multi-satellitE Retrievals for Global precipitation measurement) and reanalyses, these products can differ quite a lot from rain-gauge measurements.
- High density wind and precipitation data are required to evaluate nested modelling approaches, such as the GLObal-to-Regional ICON (GLORI) Digital Twin (Chiara Marsigli).

- At ECMWF, soil moisture observations in the tropics could help constrain (and understand) feedback processes involving the Bowen ratio and its impact on deep convection.
- With forecasting improvements, the lack of observations of gravity waves becomes more important. Perhaps vertical winds can be usefully inferred from the observations of cloud particle vertical velocities in the EarthCare mission?

Accounting for model uncertainty is essential in ensemble forecasting. This represents sub grid-scale uncertainty but also acts to project uncertainty onto large scales (e.g. via scale interactions). Can we diagnose the optimal scales for this projection, e.g. from within the MUMIP project (Hannah Christensen)?

In this session, it was shown how diagnostics can help identify deficiencies in key physical parametrizations and parametrizations of model uncertainty. We saw that a lot of care is needed in this work, with many diagnostics involving tracking, pairing with observations, compositing and conditional evaluation. Although the main focus was on fast physical processes, we also discussed slower processes in the Earth System, such as sea-ice evolution (Steffen Tietsche).

3. The tropical elephant in the forecasting room?

The question of this section aimed to elicit diagnostic evidence for the importance of the tropics in global NWP, and to provoke discussion about whether sufficient research is focused on the tropics.

Tropical waves (and their interactions with convection) shape the synoptic to planetary scale variability in the tropics, with important implications for data assimilation and predictability. There are links to tropical cyclone genesis (Philippe Peyrillé), super-position of waves can lead to extreme precipitation, and there are well-documented observed teleconnections into the extra-tropics (see Fig.11 later), for example. Not only would we like the analysed and forecast circulation to project correctly onto these waves, but also that errors and uncertainties project onto them with the correct power.

Tropical waves can be diagnosed by projecting instantaneous data onto 2D or 3D theoretical modes of the dry equations or obtained through filtering timeseries data on zonal wavenumber and frequency. Results can be sensitive to the methodology (Peter Knippertz). The projection methods have the advantage of not requiring long datasets, but the prescribed modal structures may not be so appropriate in the face of convective coupling, which is a key requirement for the development of forecasting systems. For the filtering method, the need for timeseries data is an issue for real-time monitoring but may not be a great disadvantage when it comes to systematic diagnostics. Moreover, the ability of the filtering method to allow the data to define its own (convectively coupled) modes is appealing. It means that the MJO and African Easterly Waves (AEWs) can also emerge naturally from the data. Note that filters can find waves in randomised data (Peter Knippertz), so the significance of results needs to take this into account. Fig. 6 shows that Kelvin Waves identified using a projection approach (right) tend to have coarser-scale structure and faster propagation speeds than those based on the filtering method (left) – possibly reflecting the slowing effect of convective coupling? Because of these differences, it was considered useful to include more than one identification approach in research efforts. For example, AEWs can be identified from curvature vorticity tracking (Quinton Lawton). From a community point of view, a fully accessible diagnostic tool, which can identify waves with several approaches and be applicable to ensemble forecasts, would be ideal.

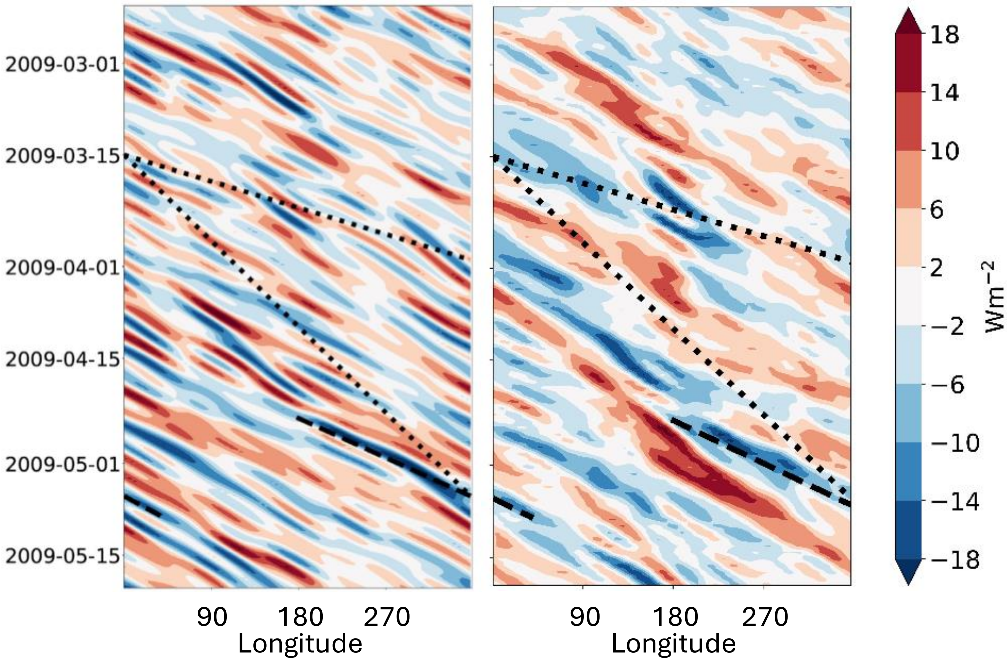


Fig. 6 *Kelvin wave power identified in out-going longwave radiation using the frequency wavenumber filtering (left) and projection onto 2D extended EOF patterns. Courtesy Peter Knippertz.*

Fig. 7 shows that, in perfect “aqua-planet” model studies, the MJO can have predictability to about 6 weeks, based on the time for the MJO index correlation to drop to 0.5 (Hyemi Kim). The effect of cooler surface temperatures over the mountains in the Maritime Continent region is to reduce predictability by 10 days. Currently predictive skill also appears to be limited by dry biases in the planetary boundary layer in the Maritime Continent region, which reduce the westward moisture flux ahead of the convectively active phase. The problem seemed reminiscent of the MCS issue discussed above. Does a lack of convective inhibition lead to the heating being placed too far ahead of the active MJO?

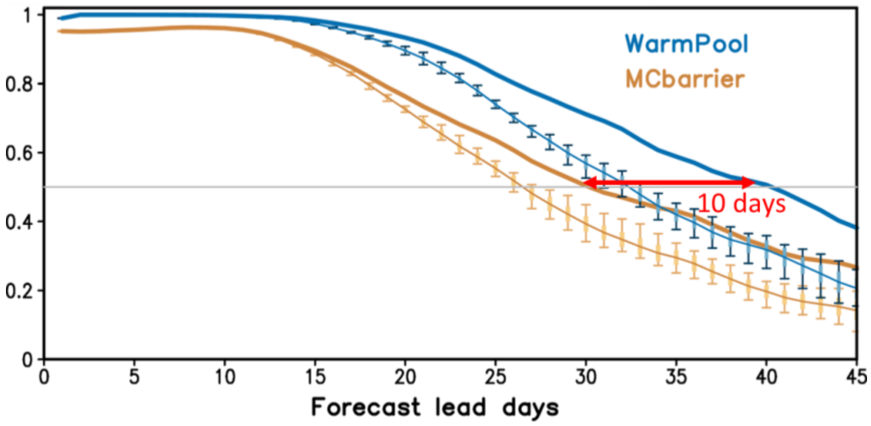


Fig. 7 *'Perfect model' anomaly correlation coefficient of the MJO index as a function of lead time in CESM2 Aquaplanet ensemble simulations. Blue: the control model. Orange: when aquaplanet SSTs are reduced using a 6.5 Kkm^{-1} lapse rate, to reflect the height of mountains in the Maritime Continent region. Bold lines show the correlation for the ensemble-mean, and thin lines show correlations for an individual ensemble member (with min, max and ± 1.0 STD indicated). Courtesy Hyemi Kim.*

Physics-dynamics coupling and interactions between physical processes get more complicated with increasing model resolution. It would be good to make more diagnostic use of process tendency data, including potential temperature and potential vorticity tendencies. Data-driven models have made big advances lately, and they show particular strengths over physics-driven models in the tropics. Their learning process tends to optimise over the physics, dynamics, and numerics altogether. Hence, they may provide a means of comparing with, and evaluating, the coupling represented in physics-driven models.

This session noted several reasons why we should invest more diagnostic effort on the tropics. It has since initiated discussions about collaborative projects focused on tropical waves.

4. Why are AI models so competitive, and are they physically consistent?

Data-driven modelling approaches to forecasting are developing at a rapid rate. For several reasons, there is a need to diagnose these models. For example, can we learn why they are so competitive (with a view to aiding physics-driven forecasting and for better understanding intrinsic predictability)? In addition, to improve confidence in data-driven forecasting (and climate projection), there is a need to investigate whether data-driven models are physically consistent. A potential answer for what we mean by "physically consistent", or at least how we might diagnose this, is whether a model displays the correct relationships between different variables. These relationships would need to be evaluated at the same resolution for the model and observations/analyses. The case study of storm Ciarán, which caused fatalities and severe damage in Europe around 2 November 2023 (Simon Driscoll), illustrates this approach to physical consistency. There is the indication that data-driven models can capture well the ubiquitous dynamical relationships, such as geostrophic balance, but maybe not the more unusual relationships (and possibly those involving diabatic aspects) such as the potential for a sting jet. A key question for future diagnostic experiments is “How well do data-driven models perform for rare/out-of-sample extreme events and in a changing climate?”

With data-driven models, it is possible to calculate the derivative of forecast error with respect to input fields, based on the full non-linear model. Fig.8 indicates that errors at the location of ex-Hurricane Ian (black box) are sensitive to initial surface temparatures in the Caribbean/Gulf of Mexico, close to where the Hurricane formed (Uroš Perkan). This result is reminiscent of ensemble forecast sensitivities for Hurricane Humberto in September 2019, based on the IFS. Such diagnostics could provide useful insight into predictability, and observation impact, including at lead times beyond the linear regime.

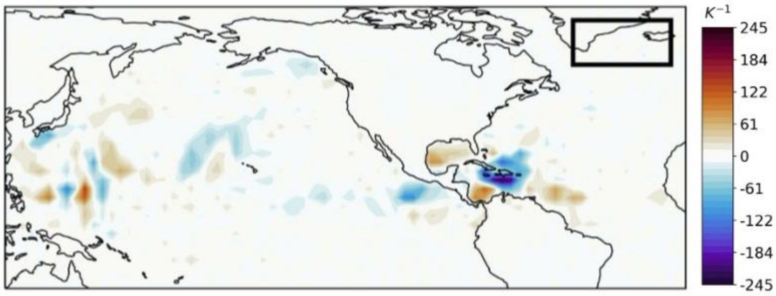


Fig. 8 *Sensitivity of the normalised 10m wind error in the indicated box at lead time 12 days to initial surface temperatures on 23 September 2022. Courtesy Uroš Perkan.*

There is also the potential to use relaxation techniques for comparing teleconnections in data-driven and physics-driven models with reality (Uroš Perkan).

We also discussed how users are adding value to forecasts through application of artificial intelligence. The energy sector is already making use of (e.g.) deterministic forecasts made using ECMWF’s Artificial Intelligence Forecasting System (AIFS). Statistical post-processing of these forecasts, based on ERA5, can produce 100m winds which out-perform the IFS in terms of skill (Isla Finney). A key issue is forecast inconsistency (the degree of disagreement between forecasts for a given validity date but decreasing lead-time). With increasing numbers of forecast systems available (data-driven and physics-driven), information and diagnostics that help the user decide which system might be best for a given application are becoming increasingly important.

This session helped inform the wider diagnostic community about the potential of data-driven approaches, and discussions suggested avenues for diagnostic development. Work at ECMWF, for example, is beginning to diagnose the important “physical consistency” question.

5. Does resolution matter in global NWP?

Kilometre-scale limited area models (LAMs) improve many aspects at short time ranges, such as the diurnal cycle, extreme precipitation, tropical cyclone intensification, and orographic and coastal effects (Claudio Sanchez). A key question is whether these improvements would lead to better longer-range forecasts of the large-scale circulation in km-scale global models? This question has been explored with the Met Office's "K-Scale hierarchy", which drives a Cyclic Tropical Channel model at 5km (CTC5) or a LAM at 2km with boundary conditions from a global model (GLM) at 10km, and has the choice of full or reduced parametrization suites; in particular the choice of whether to parametrize convection or let the model attempt to resolve convection. Results suggest that the reduced parametrization (regardless of resolution) leads to a better diurnal cycle in precipitation, and a better African Easterly Jet and associated MCSs. It also extends the shallower "-5/3" spectrum for kinetic energy into the mesoscales, making the model more "chaotic". Indeed, in twin experiments, the choice of parametrization appears to have a larger impact on Kelvin wave uncertainty, than does the resolution, Fig.9. A key question is whether we need resolution to justify turning-off physical parametrizations, or can forecast reliability be obtained at lower resolution with physical parametrizations turned on and the inclusion of improved model uncertainty?

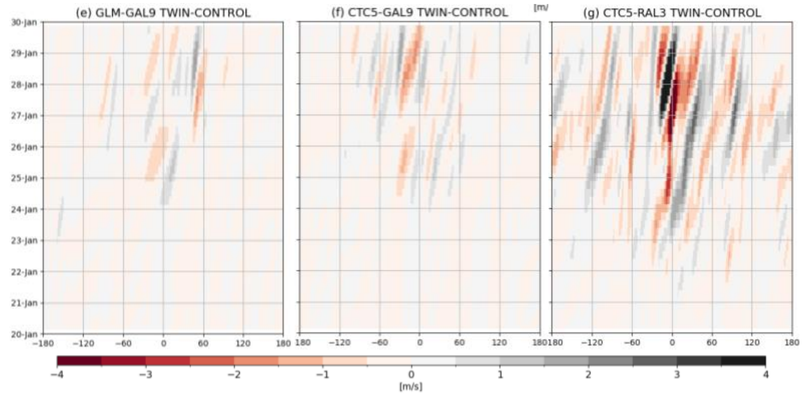


Fig.9 *Difference in 200hPa zonal wind associated with equatorial Kelvin Waves in twin forecast experiments with (left) GLM with full parametrization suite, (centre) CTC5 model with full parametrization suite and (right) CTC5 model with reduced parametrization, including no convective parametrization. Kelvin waves were identified through projection onto 2D dry theoretical modes.*

Global power spectra diagnostics can show the spatial-temporal evolution of forecast uncertainty and reliability, and the impacts of model and observational changes, Fig. 10 (Mark Rodwell). By separating scales, power spectra also allow a fairer comparison of forecast systems run at different resolutions. For regions which are non-global or even non-zonally-complete, other approaches such as a diffusion-based filter can be used (Matthias Aengenheyster). Additional measures of forecast performance can also be diagnosed in a scale-dependent way, such as forecast inconsistency and precipitation errors (Zhao Bin).

The question of the need for higher resolution in global modelling has become even more relevant with the evident success of low-resolution data-driven forecasts. While power spectra indicate that there is little predictability at small scales in general, is the lack of interactions with these small scales also unimportant for larger-scale predictive skill? Further diagnostic work aimed at answering these questions will be important.

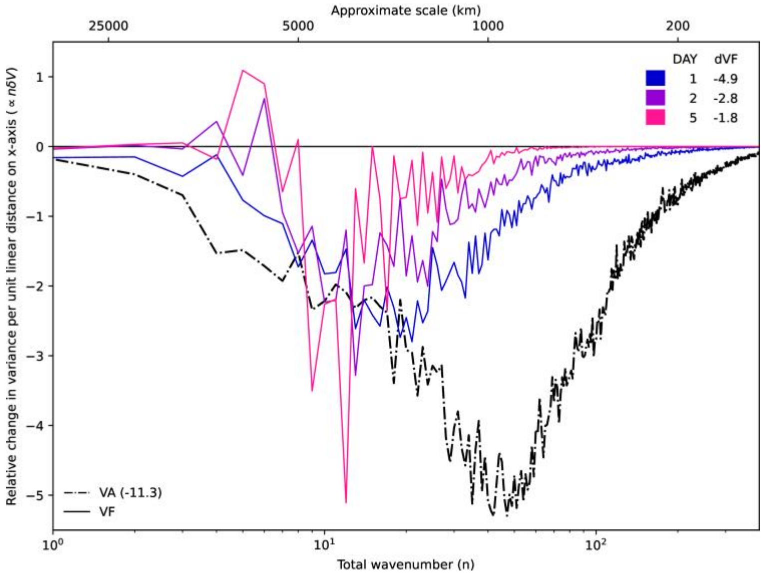


Fig. 10 *The impact of 30,000 new Radio Occultation observations* each analysis cycle on the ensemble variance of the analysis (VA) and forecast (VF), shown as contributions to the total percentage change. *Based on data from the IROWG ROMEX initiative for September 2022. Experiments by Katrin Lonitz.*

6. The importance of community datasets and intercomparison

It is promising to see that there are open-source datasets and toolboxes for the community to use in analysing weather and climate data. Two such diagnostic packages were presented, showcasing a variety of tools and visual output options, followed by discussion of ERA5 and ERA6 reanalyses.

The MJO teleconnections package (Fig. 11, Chaim Garfinkel) is run through a graphical user interface (GUI), where relationships based on the STRIPES index (Sensitivity To the Remote Influence of Periodic EventS), pattern correlation coefficient, extra-tropical cyclone activity, and surface air temperature can all be implemented, with bootstrapping providing statistical significance testing. The Ensemble Museums tool (Mio Matsueda) displays archived graphical products using data from TIGGE (up to 2 weeks ahead), S2S (up to 2 months ahead), and C3S (up to 6 months ahead) and is helpful in finding past forecast cases. The TIGGE dataset consists of medium-range ensemble forecast data from 13 global NWP centres, with 16 forecast products from October 2006 to present, including Z500, MJO, winter weather regime forecast, probabilistic forecast of severe weather events.

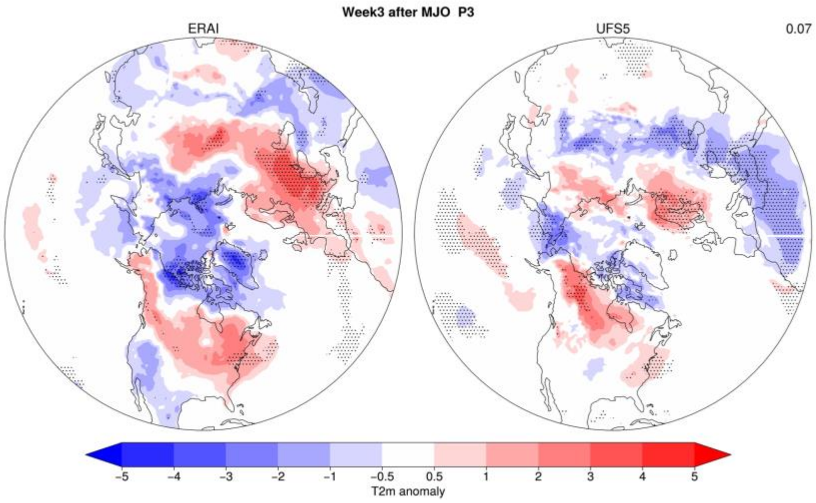
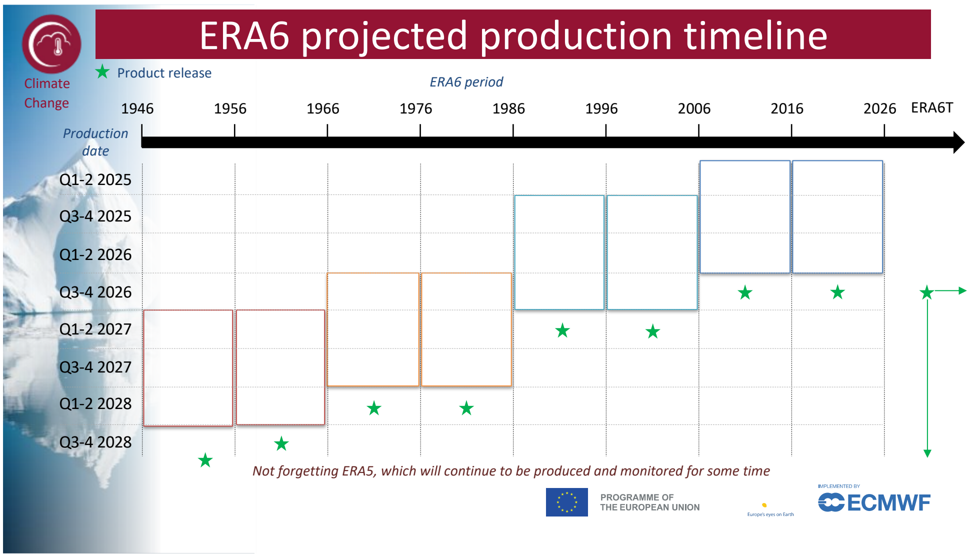


Fig.11 *MJO teleconnections tool showing the relationship with surface temperature 3 weeks after phase 3 of the MJO, based on (left) ERA-Interim and (right) NOAA Unified Forecast System prototype 5 (UFS5). Courtesy Cristiana Stan.*

ERA5 is a valuable tool for diagnostics, used as a 'ground truth' for many phenomena at a variety of temporal and spatial scales. ECMWF routinely monitors ERA5 using default indicators, such as observation departures, analysis increments, and anomalies from climatology. ERA6 will have 14 km (TCo799) resolution and an 11-member analysis ensemble (28 km). Production starts early 2025, with the current plan for the first release of 2006-present data at the end of 2026.



ERA6 projected production timeline

Climate Change

★ Product release

ERA6 period

Not forgetting ERA5, which will continue to be produced and monitored for some time

PROGRAMME OF THE EUROPEAN UNION

ECMWF

Fig.12 *Projected production and release timeline of ERA6. Courtesy Alison Cobb.*

The community is keen on having easier access to the diverse range of observational data for verification. A first step may be an inventory, with the ideal solution having all data in one place. Discussions are taking place within WMO working groups.

7. Summary

The workshop was very successful, with a broad range of experts participating from around the world (Fig. 13). It is highly recommended to visit the superb talks and posters on the workshop portal (https://events.ecmwf.int/event/383/). Here, we list a few key conclusions of the workshop.

Diagnostics of intrinsic predictability suggest that forecast predictions of day-to-day weather could be improved by about 4-5 days. A pre-requisite in progress towards this goal would be a reduction in analysis uncertainty (standard deviation) by a factor of about 10. For physics-driven models, such predictive skill would require improvements to the representation of many interacting processes. One particularly interesting consensus that emerged from the workshop centred on the importance of the timing of deep convection. This seems to affect the development of mesoscale convective systems, the impacts of warm conveyor belts on the upper troposphere, interactions with tropical waves, and the propagation of the MJO. Moreover, uncertainties in this timing appear to be associated with the fastest rates of forecast uncertainty growth. Diagnostics that target this aspect from different angles would be particularly useful for forecast system development (some have already been discussed here).

Data-driven models display incredible skill, and there is a need for diagnostics that help us understand this success. Diagnostics of their physical consistency will be important to build trust in their predictions of “out-of-sample” events. Are they successful despite currently having low resolution, or does their success imply resolution is less important than previous thought? Interestingly, the differentiability of non-linear data-driven models can provide a new perspective for predictability studies.

There is clear strategic need to develop a diagnostic tool for convectively-coupled tropical waves. Such a tool would be useful throughout the forecasting process: from the evaluation of observational information to ensemble data assimilation, global ensemble forecasting and, via ocean coupling, to seasonal timescales.

Re-analysis is a source of “ground truth” for many diagnostic studies and the community will greatly appreciate the release of ERA6.

Finally, we should be mindful of the vital step from diagnosing an error, to fixing it! This is not as trivial as it sounds. The nature and magnitude of forecast system errors evolves with every system update, and with the introduction of new observations. This means that diagnostics also need to evolve, while remaining understandable and straightforward to operate.

Fig. 13 *Participants taking part in the workshop.*