Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

927

Forecast verification using Information and Noise

Massimo Bonavita and Alan J. Geer

16 May 2025

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

Abstract

Numerical Weather Prediction (NWP) Centres evaluate forecast quality using statistical assessments of error and skill, commonly referred to as scores. Traditional forecast verification relies on metrics such as Root Mean Squared Error (RMSE), Anomaly Correlation Coefficient (ACC), Brier Score, etc., which compare forecast performance relative to baseline models. However, these measures mix bias effects and forecast uncertainty, complicating direct comparisons between deterministic models, ensemble averaging approaches, and machine learning-based forecasts. A clear separation between intrinsic forecast skill and post-processing enhancements, such as calibration, is essential for accurately assessing the predictive capability of a forecast system.

In this work we take forecast reliability and resolution as the fundamental attributes characterising forecast performance, with resolution representing the true predictive capability of a system—its ability to distinguish among observed events. Recent work by Feng, Toth, Zhang and Peña, 2024, introduced Information and Noise as new metrics designed to provide an unambiguous assessment of statistical resolution. This study aims to introduce these novel scores in an accessible manner, relating them to traditional verification metrics, and tackles some of the limitations of the original formulation. Additionally, we demonstrate their practical implementation for routine forecast verification in an operational NWP environment, and provide examples of their use in the standard NWP research workflow. Examples of application of these new verification metrics to ensemble forecasting and to machine learning forecast models are also provided.

Plain Language Summary

Weather forecasting centres use statistical verification to check how accurate predictions are. Traditional verification metrics give estimates of total error, which is important for end users of forecast products, but does not allow to evaluate the intrinsic information content of the forecasts and makes it hard to fairly evaluate different forecasting approaches, such as single models, averages of multiple forecasts, and AI-based predictions.

In this study we adopt the view that forecast reliability (how often predictions are correct) and resolution (how well a forecast distinguishes between different possible weather outcomes) are the two key attributes for assessing forecast quality. A new set of metrics, called Information and Noise, has been introduced to measure resolution more clearly. The study explains these new metrics in simple terms, connects them to traditional methods, and addresses their previous limitations.

Additionally, it shows how these new scores can be used in daily forecasting operations and research, with examples including ensemble forecasts (which use multiple predictions) and AI-driven weather models.

1. Introduction

Numerical weather prediction centres measure the quality of their forecasts using statistical assessments of error and skill, typically known as 'scores'. For deterministic forecasts (also known as categorical or non-ensemble forecasting), these include the root mean squared error (RMSE) and anomaly correlation coefficient (ACC, e.g. Jolliffe and Stephenson, 2003). A skill score typically measures the error relative to a baseline, such as the quality of a forecast based on climatology. Scores are used to compare the quality of forecasts between NWP centres (e.g. Haiden et al., 2023) and to assess the impact of proposed new developments, such as activating new observations (e.g. Healy et al., 2024).

The correspondence between forecast scores and the true skill of a forecast is not exact. For example, RMSE is sensitive to bias as well as the error standard deviation. Hence, in areas where the error growth is relatively low, such as the tropics, changes in bias that are essentially irrelevant can overwhelm other signals. It is also hard to compare the outputs of traditional deterministic models with those based on ensemble averaging or on unperturbed ML-based models, since these are typically smoothed relative to reality, and most scores are sensitive to changes in forecast activity as well as changes in true skill (e.g. "double penalty effect", Hoffman et al., 1995; Ebert et al., 2013). Hence it is important to separate the true underlying skill of a forecast system from the aspects that can be improved by smoothing, filtering, bias correction and so on, activities collectively known as calibration and which can be performed in a post-processing step.

Statistical resolution and statistical reliability are two forecast attributes that completely characterise forecast performance (e.g. Jolliffe and Stephenson, 2003; Toth, Talagrand and Zhu, 2005, TTZ05 in the following; Candille and Talagrand, 2006; Feng, Toth, Zhang and Peña, 2024, FTZP24 in the following). Both forecast attributes are important for forecast end users. However, lack of statistical reliability can be corrected a-posteriori to the same degree that it can be diagnosed (under the hypothesis of stationarity of the underlying statistics). On the other hand, statistical resolution measures the true or intrinsic predictive capability of a forecast system (Wilks, 2011), i.e. the ability of the forecast system to sort observed events into groups that are different from one another, and it cannot be improved with post-processing techniques.

The more widely used forecast skill measures in the operational NWP community, e.g. Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Brier Score (BS), Ranked Probability Score (RPS), Continuous Ranked Probability Score (CRPS), etc., mix the two forecast attributes described above. While decompositions into resolution and reliability components are available for the probabilistic scores (e.g., BS decomposition, Murphy, 1973), these decompositions are not routinely used in operational practice. Additionally, they are limited to probabilistic forecasts. FTZP24 have proposed the skill measures Forecast Information and Noise to provide an unambiguous diagnostic of the statistical resolution of a forecast system. The aim of this work is: (a) to provide a more accessible introduction to the new scores, including elucidating their relationship to forecast skill scores that are widely used

in the operational NWP community; (b) to show how it is possible to practically implement the new verification framework for routine forecast verification in an operational NWP environment (ECMWF); (c) to show examples illustrating the different qualities of the deterministic and ensemble forecasts at ECMWF; (d) to show how the scores can be applied in the standard NWP Research and Development workflow; (e) to give an example of their use to evaluate new Machine Learning models.

2. A geometric view of forecast scores

Typically, deterministic forecast scores are generated from spatial fields over a defined area such as the northern hemisphere extratropics. These statistics are computed for a single variable, a single level in the vertical, and a single forecast validity time. In this approach, each of the $i = 1$ to n locations in the spatial field can be assigned a position in a vector. If x_i represents the geophysical variable at each location, we could define a vector $\mathbf{x}''$ to represent them all: $\mathbf{x}'' = (x_1, x_2, \ldots x_n)$. This geometric view treats each location i as being represented by a different dimension within an n-dimensional space (technically, an orthonormal basis). As will shortly become clear (and see also appendix A) it is more convenient to apply a scaling when creating the vector, and we will define this vector $\mathbf{x}$ (rather than $\mathbf{x}''$)

$$\mathbf{x} = (wx_1, wx_2, \ldots wx_n). \qquad (1)$$

Here, w is the scaling factor, which we will set to $\frac{1}{\sqrt{n}}$ initially. With this scaling, distances between points in the n-dimensional space are equivalent, and mathematically equal, to Root Mean Squared Differences (and Root Mean Squared Error if one of the points is the verifying "true" state). This geometric approach will provide an intuitive visual explanation of the relationships between different scores, and moreover, it is very useful in explaining the meaning of the new noise and information scores of FTZP24.

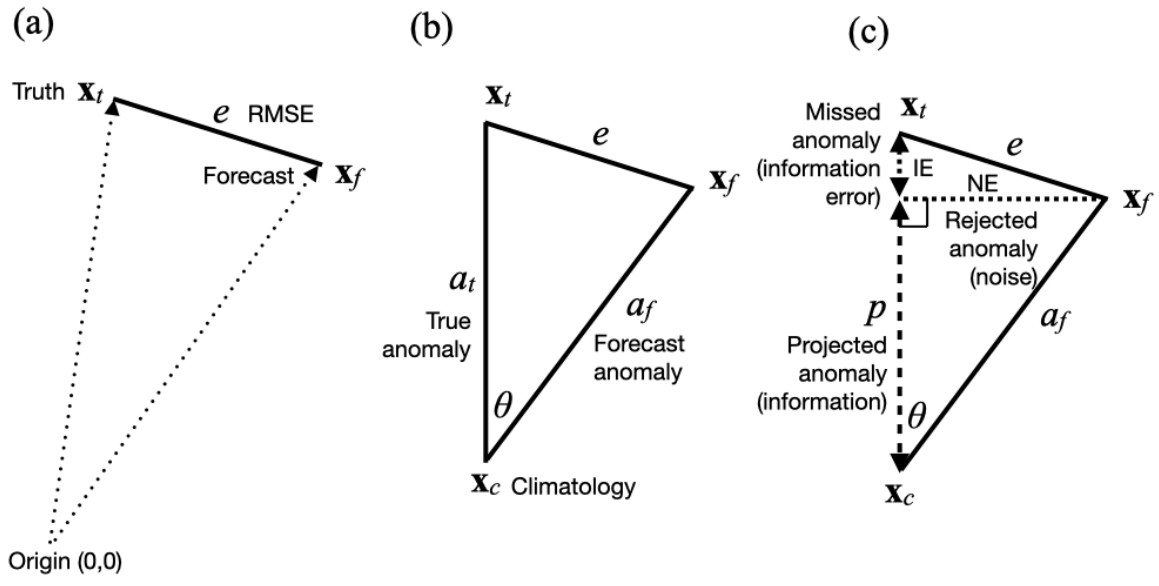

Figure 1: Geometrical decomposition of forecast errors and anomalies. See text for details.

Panel (a) in Figure 1 illustrates the verification of a forecast in a 2-dimensional space after applying the scaling factor w. The vectors $\mathbf{x_f}$ and $\mathbf{x_t}$ represent the forecast and the validating 'truth'. The vector $\mathbf{x_f} - \mathbf{x_t}$ is the forecast error. To summarise the size of this error in a single number, we can use the length of this vector, in other words the length of the solid line on Figure 1a. To compute this, we can make use of the vector dot product, which is defined between two vectors $\mathbf{a}$ and $\mathbf{b}$ as:

$$\mathbf{a} \cdot \mathbf{b} := \sum_{i=1}^{n} a_i b_i \qquad (2)$$

Here, a_i and b_i represent the magnitude of the vector in each of the specific n directions. The length of any vector is computed, equivalently to Pythagoras' theorem extended to arbitrary dimensions, as the square root of the dot product of a vector with itself. The length of vector $\mathbf{a}$ is just $\sqrt{\mathbf{a} \cdot \mathbf{a}}$. Note that we prefer to denote length in this way, rather than using the Euclidean norm notation, in order to simplify the presentation. For the forecast error, its length is a scalar:

$$e = \sqrt{(\mathbf{x_f} - \mathbf{x_t}) \cdot (\mathbf{x_f} - \mathbf{x_t})} \qquad (3)$$

When expanded using equations (1) and (2) this recovers the formula for the RMSE between forecast and truth:

$$\mathrm{RMSE}(\mathbf{x_f}, \mathbf{x_t}) = e = \frac{\mathbf{1}}{\sqrt{n}} \sqrt{\sum_{i=1}^{n} (x_{f,i} - x_{t,i})^2} \qquad (4)$$

Hence the RMSE is a natural measure of the error in a forecast because it is the length of the error vector. Note that this includes the standardisation for the number of dimensions using the weighting factor $w = \frac{1}{\sqrt{n}}$ introduced earlier (because the length of the vector (1,1,1….1) is $\sqrt{n}$, the length standardisation needs to change according to the number of dimensions).

We can now introduce another vector $\mathbf{x_c}$ representing climatology, which is typically a monthly mean of the relevant analysis field computed over many years. A typical skill score evaluates how much better the forecast is compared to such a climatology (e.g. Murphy, 1988). In the geometric view, shown in Figure 1b, the end points of the vectors for climatology, truth and forecast form a triangle. The other sides of the triangle are the vectors $\mathbf{x_f} - \mathbf{x_c}$, which is the forecast anomaly, and $\mathbf{x_t} - \mathbf{x_c}$, which is the true anomaly, relative to climatology. The lengths of these sides of the triangle are $a_f := \sqrt{(\mathbf{x_f} - \mathbf{x_c}) \cdot (\mathbf{x_f} - \mathbf{x_c})}$ and $a_t := \sqrt{(\mathbf{x_t} - \mathbf{x_c}) \cdot (\mathbf{x_t} - \mathbf{x_c})}$**.** These quantify the size of the anomalies of the forecast and of the truth. The cosine of the angle (θ) between the forecast anomaly and true anomaly gives the anomaly correlation. This can be seen by applying the dot product cosine identity

$$\mathbf{a} \cdot \mathbf{b} := \sqrt{\mathbf{a} \cdot \mathbf{a}}\,\sqrt{\mathbf{b} \cdot \mathbf{b}}\cos(\theta) \qquad (5)$$

to the forecast anomaly and the true anomaly vectors, and solving for $\cos(\theta)$, to define what we will call the RMS-derived Anomaly Correlation Coefficient (RACC):

$$\mathrm{RACC} := \cos(\theta) = \frac{(\mathbf{x_f}-\mathbf{x_c})\cdot(\mathbf{x_t}-\mathbf{x_c})}{a_f a_t} \qquad (6)$$

These quantities are all linked by the vector version of the cosine rule, which relates the lengths of the sides of a triangle to the angle between two of the sides. Hence

$$e^2 = a_f^2 + a_t^2 - 2a_f a_t\cos(\theta) = a_f^2 + a_t^2 - 2a_f a_t\mathrm{RACC} \qquad (7)$$

This is nearly equivalent to the decomposition of the mean square error into activity and anomaly correlation that is familiar from works such as Murphy and Epstein (1989, compare to their equation 9). If we can define RMS-like activity of the forecast and verification (truth) as RMSDAF = a_f and RMSDAV = a_t, then equivalently, but with alternative notation:

$$\mathrm{RMSE}^2 = \mathrm{RMSDAF}^2 + \mathrm{RMSDAV}^2 - 2\,\mathrm{RMSDAF}\,\mathrm{RMSDAV}\,\mathrm{RACC} \qquad (8)$$

Note that in practical forecast verification, two additional corrections are included. The first, as included in Murphy and Epstein's decomposition, is that mean anomalies are subtracted before computing the lengths of the sides of the triangle, so that the triangle is composed of the sample standard deviation equivalents of the RMS measures we have used here, i.e. standard deviation of error and anomaly SDE, SDAV and SDAF and the standard anomaly correlation coefficient ACC, also the cosine of the angle between the mean-corrected anomalies. The second correction, not included in the aforementioned work, is that area weighting is applied. This recognises that atmospheric fields are usually provided on a regular lat-lon grid, so the statistics would give excessive weight to polar regions if computed naively. The results we show later include these two corrections. The relevant extensions and proofs are presented in appendix A.

The novelty and insight of FTZP24 is to consider the projection of the forecast anomaly $\mathbf{x_f} - \mathbf{x_c}$, onto the true anomaly $\mathbf{x_t} - \mathbf{x_c}$, as shown in Figure 1c. The length of the projected anomaly is denoted p and is known simply as Information. When compared to the length of the true anomaly, this defines their newly proposed measure of Forecast Information, FI, as the ratio of the two:

$$\mathrm{FI} = p/a_t \qquad (9)$$

When FI=1, the anomaly has been forecast perfectly, and hence the forecast was perfectly informative. When FI=0, the forecast contained no information on the observed anomaly and thus was not informative. Note that FTZP24 defined FI as the square of the quantities we use here ($p^2/{a_t}^2$) so they were working in the sense of variances. But we prefer to keep

consistency with existing verification statistics such as the RMSE, standard deviation and ACC, so we suggest that FI should be defined our way in future. We also note that Murphy and Epstein (1989) argued that the square of ACC was a fairer measure of potential forecast skill, but this suggestion did not catch on in the verification literature, and the use of measures of forecast skill in the sense of standard deviation remains standard practice.

The length p is obtained from the length of the forecast anomaly and $cos(\theta)$, by scalar projection, so that

$$p = a_f \, cos(\theta) = a_f \, RACC \qquad (10)$$

Hence, forecast information FI is closely related to the anomaly correlation, but with an alternative normalisation (compare to equation (6)):

$$\mathrm{FI} = \frac{a_f}{a_t}\mathrm{RACC} = \frac{(\mathbf{x_f}-\mathbf{x_c})\cdot(\mathbf{x_t}-\mathbf{x_c})}{a_t^2} \qquad (11)$$

The alternative normalisation makes FI sensitive to changes in the forecast activity. Specifically, FI penalises forecasts that show levels of activity smaller that the observed activity, e.g. unrealistically smooth forecasts. In contrast, anomaly correlation is unaffected by forecast activity. RMSE can also be artificially reduced (see e.g., equation 8) just by reducing the forecast activity, without making any improvement in underlying forecast skill. Appendix B examines some edge cases with FI and explains why we have chosen to define it as described above. In particular, note that the angle theta can be larger than 90 degrees, a situation which leads to negative anomaly correlation through the cosine in equation 6. We also allow negative values of p in equation 10 so that FI can go negative in the same circumstances, which makes our definition of FI (equation 9, 11) equal in magnitude to the square root of the FTZP24 equivalent, but with a varying sign. These considerations are important where anomaly correlation and the forecast information approach or go below zero, such as at longer forecast ranges.

A closely related set of scores is obtained by considering the upper right-angled triangle in Figure 1c. This provides the orthogonal decomposition of the forecast error along the direction of the true anomaly and its orthogonal complement. The size (length) of the error in the direction of the true anomaly will be called the Information Error (IE) and the size of the error orthogonal to the true anomaly will be called the Noise Error (NE). Specifically:

$$\mathrm{IE} = abs(1 - FI)\, a_t = abs\left(a_t - a_f RACC\right) \qquad (12)$$

Note that FI is not bounded by 1 and in case of forecast anomalies larger than observed anomalies it can be larger than 1. Hence the abs() is practically required to avoid negative lengths but, more fundamentally, it is a consequence of the definition of IE as an error measure (see appendix B). The aim is that IE and NE, as an orthogonal decomposition of RMSE, obey Pythagoras' theorem, hence

$$\mathrm{RMSE}^2 = \mathrm{IE}^2 + \mathrm{NE}^2 \qquad (13)$$

NE can be obtained through this identity from RMSE and IE. However, we can also consider the **vector projection** of the forecast anomaly on the true anomaly, which is FI $(\mathbf{x_t} - \mathbf{x_c})$. Then the vector noise is the corresponding rejection:

$$\mathbf{r} = (\mathbf{x_f} - \mathbf{x_c}) - \mathrm{FI}\,(\mathbf{x_t} - \mathbf{x_c}) \qquad (14)$$

And the normalised length of this vector defines NE as

$$\mathrm{NE} = \frac{\sqrt{\mathbf{r}\cdot\mathbf{r}}}{\sqrt{n}} \qquad (15)$$

Note that this is different from the definition of Forecast Noise, FN, by FTZP24, which is equivalent to $\mathrm{FN} = (\mathbf{r}\cdot\mathbf{r})/a_t^2$. FTZP24 normalise by the true anomaly variance for consistency with forecast information FI. Instead, our definition provides a consistent orthogonal decomposition of the RMSE into information error (IE) and noise error (NE).

We can now recap Figure 1, starting with panel a, which shows the geometric interpretation of RMSE, for a forecast validated against a "truth", as the standardised length of the error vector. Panel b shows that the forecast, truth and climatology define a triangle (in a n-dimensional state space) in which the anomaly correlation is the cosine of the angle between the true anomaly and the forecast anomaly. As noted by Murphy (1988), this makes anomaly correlation insensitive to the amount of activity in the forecast (as well as to bias in the forecast) so he argued it is a measure of potential rather than actual skill. Murphy and Epstein (1989) hence also noted that any interpretation of anomaly correlation is incomplete without examination of additional terms such as activity and mean error. This point can be clearly understood from Figure 1b. Further, it is obvious that, if anomaly correlation stays constant, RMSE can still be minimised by making the error vector perpendicular to the forecast anomaly, as shown in Figure 2. This illustrates a well-known problem with RMSE: that it can typically be improved by decreasing the forecast activity, potentially favouring forecast systems with an unrealistic representation of the atmosphere. However, the reduction in total forecast error described above will be the result of a decrease in Noise Error and a concomitant increase in Information Error, i.e. the information content of the damped forecast is reduced.

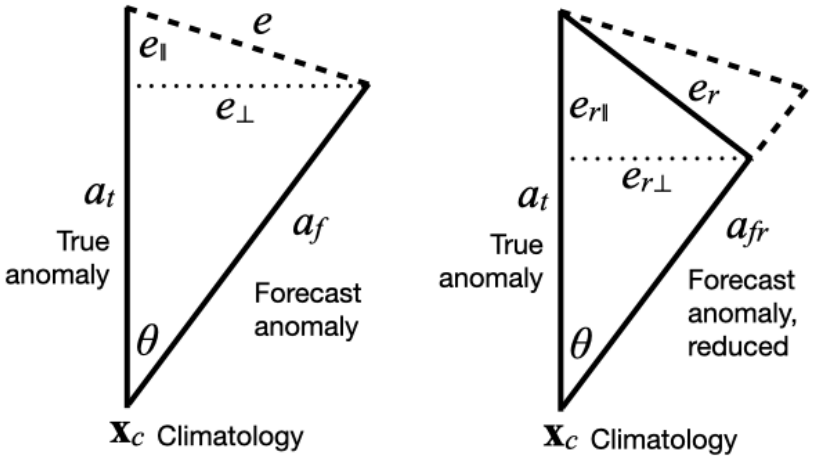

Figure 2: Geometrical illustration of how RMSE can be reduced by damping forecast activity, reducing forecast noise, but at the cost of larger information error (equivalently, less information).

Figure 1c shows the FTZP24 decomposition of the forecast anomaly along and perpendicular to the true anomaly, which measures information and noise in the forecast. FTZP24 showed that although ensemble mean forecasts have apparent advantages in terms of higher anomaly correlation and lower RMSE, this is achieved primarily by reducing noise rather than by adding any new information. Indeed, the information in the ensemble mean is smaller than that of a deterministic forecast, apparently due to the stochastic perturbations used to create the ensemble members. Such an analysis illustrates the power of the new approach.

We have seen that forecast information FI is very similar to anomaly correlation except for its normalisation. However, anomaly correlation is by construction bounded by 1, but FI can generate values larger than 1 if the forecast anomaly is larger than the observed anomaly. We have already seen this issue in practical verification on some days, in the early forecast range. Once the scores are aggregated across multiple days, this can confuse the interpretation: for example, a perfect aggregated score of 1 could be achieved by a mix of forecasts with under- and over-shooting FI. Hence, we prefer to consider the upper right-angled triangle on Figure 1c, which is the decomposition of the RMSE into Information Error (IE) and Noise Error (NE), rather than the forecast information and noise of FTZP24. With this decomposition it is unambiguous that a perfect forecast would have both zero IE and zero NE, with zero being a true lower bound. It is then possible to understand any change in the RMSE in terms of noise and information error. Crucially, a reduction in RMSE only indicates an improved forecast if it does not increase the information error. Monitoring of IE would discourage the possible gaming of RMSE by a reduction in forecast activity and help put on a quantitative basis the effects of changes in forecast activity on forecast skill. In the orthogonal direction, any reduction in RMSE caused by reduced forecast noise may be a welcome improvement, though it does not increase the underlying information content (or statistical resolution; Wilks, 2011) of the forecast, and may reduce the forecast activity below realistic levels.

Further details on the practical implementation of the skill and error measures in the ECMWF verification package Iver (Geer, 2016) are provided in Appendix A, along with full generalisation of the mathematical presentation to include standard-deviation type statistics (SDE, SDAV, SDAF, ACC) and area weighting. It is worth emphasising again that in this introduction, we have used the simpler decomposition triangle of RMSE into RMSDAV and RMSDAV, with RACC the cosine of the angle between the latter two, where these are the versions of the activity and the anomaly correlation formulas computed without subtraction of the mean. But in standard practice, the mean is subtracted, so that the decomposition triangle is properly STDE, SDAV and SDAF, with ACC the cosine of the angle between the activities. In the rest of this report, the new verification statistics Forecast Information (FI), Information Error (IE) and Noise Error (NE) are computed, for consistency, with the subtraction of the mean (as well as applying area weighting and averaging across samples) so they are part of the STDE decomposition, not the RMSE decomposition.

3. Medium Range ensemble forecasts

The ECMWF medium-range ensemble prediction system (ENS) is an ensemble of 51 forecasts with a horizontal resolution of around 9 km. It comprises one control (unperturbed) forecast (CF) plus 50 perturbed forecasts (PF) each run from perturbed initial conditions and model physics (Owens and Hewson, 2018, Chap. 5). The perturbations in the initial conditions and model physics are aimed at sampling uncertainties in the initial conditions and model errors.

Below we show results documenting the application of the Information Error and Noise Error measures to the forecasts of the ENS control (unperturbed) member, the ENS ensemble forecast mean (EM) and perturbed members 1, 10, 20, 30, 40, 50 (PF1, PF10, PF20, PF30, PF40, PF50) for 500 hPa geopotential (Z500). Similar results are available for other forecast quantities (e.g., 2-metre temperature (T2m), mean sea level pressure (MSLP), but are not shown for conciseness.

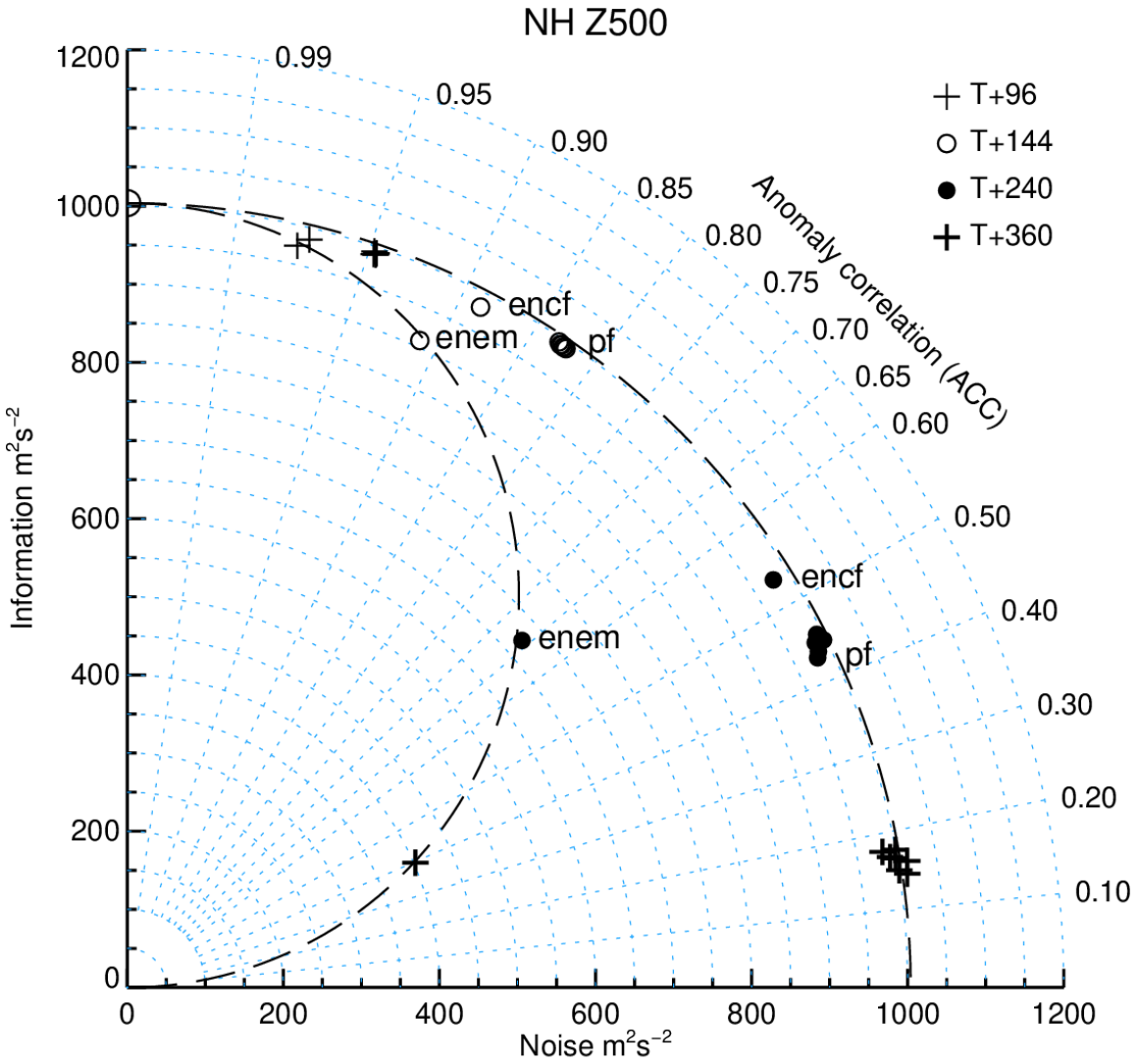

Figure 3: Information, Noise, ACC diagram for the unperturbed (control) forecast (encf), for the ensemble mean (enem) and the perturbed forecasts (pf). Information and noise increase along the y and x axes respectively (they are an orthogonal decomposition of the forecast anomaly). Distance along the radials represents forecast anomaly, with the true (analysed) anomaly represented by the dashed quadrant line. The dashed semi-circle indicates the position of lowest STDE for a given anomaly correlation (which is where STDE and SDAF are orthogonal as shown in Figure 2). The semi-circle on the y axis represents the true (analysed) anomaly, in other words a perfect forecast by our definition. The straight-line distance from this point to any of the forecasts is the relevant forecast error standard deviation, STDE.

Figure 3 shows the results for NH 500hPa geopotential height in the orthogonal space defined by information and noise following FTZP24 and our earlier theoretical discussion. For the 10 day (T+240h) forecast, the ensemble mean has anomaly correlation above 0.7, which appears superior to the unperturbed forecast at around 0.6 (and the straight line distance from the true anomaly, the error standard deviation, is also shorter). However, the unperturbed forecast information of 626 compared to 549 (in m^2s^{-2}) suggests that the unperturbed forecast has

significantly higher statistical resolution, i.e. higher fundamental forecast skill than the ensemble mean. The apparent advantage of the ensemble mean in terms of ACC and STDE is obtained by reducing activity (distance from the origin in the radial direction) and noise (distance along the x axis). The ensemble members have similar information to the ensemble mean (lower than the unperturbed forecast) but more realistic levels of activity. For the 15 day (T+360h) forecast, the unperturbed forecast seems to be no better than the ensemble members, and we can speculate that, if further extended in time, all the forecasts will continue to follow the dashed line to a point where they continue to have realistic activity but no information. In contrast, the trajectory of the ensemble mean closely follows the location of minimum standard deviation of error, indicated by the dashed semi-circle. This shows that, if the aim of the ensemble forecast is to reduce the STDE to its theoretical minimum for any given ACC (the dashed semi-circle), then the ECMWF ensemble mean achieves this almost perfectly. It seems highly likely that if the ensemble mean forecast were further extended, it would follow this line spiralling to the origin, to a point where it contains neither information nor forecast activity.

Further plots show a more familiar breakdown of the results as a function of forecast time. Qualitatively, the plots are also coherent and similar to the equivalent plots presented in FTZP24 (e.g., their Fig. 8, panels b, c). Specifically:

a) In terms of RMSE (Figure 4, top row) the ensemble forecast mean (EM) has lower error than the ensemble control unperturbed forecast (CF) which has, on average, lower error than any perturbed forecast;
b) The Information Error (IE; Figure 5, top row) of the unperturbed forecast is smaller, on average, than those of the EM and perturbed forecasts: for Z500 in the northern extra-tropics the difference is approx. 18-24 hours for lead times from 5 to 10 days;
c) The Information Error (IE; Figure 5, top row) of the ensemble forecast mean (EM) is, on average, indistinguishable from those of any perturbed forecast members. This supports results from FTZP24, their Figure 6;
d) The Noise Error (NE; Figure 5, bottom row), which is the projection of total forecast error onto directions orthogonal to the true anomaly (verifying analysis) is largest for perturbed forecasts and smallest for EM, with the unperturbed forecast in the middle.

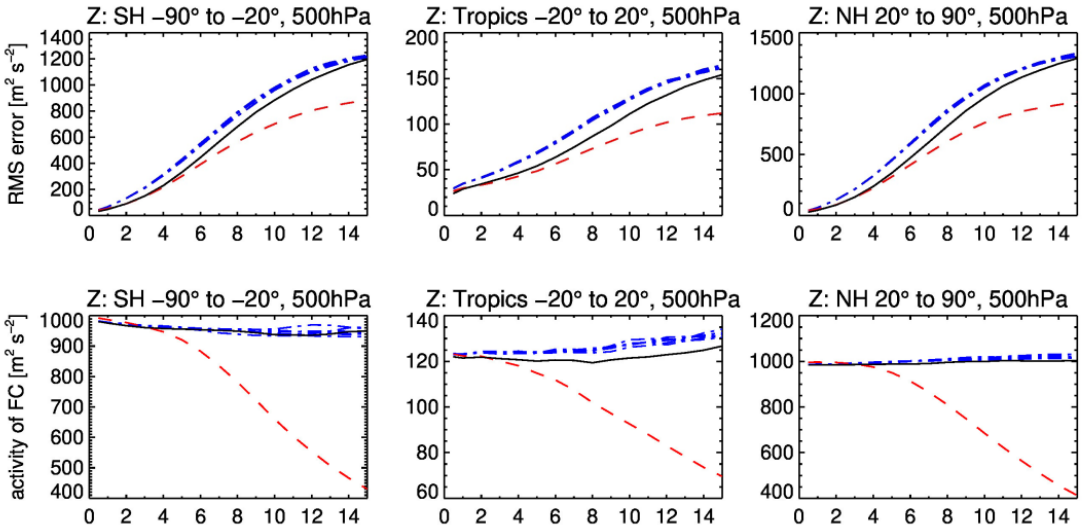

Figure 4: Top row: Evolution of the ENS Prediction System Forecast RMSE of 500 hPa geopotential (Z) for the unperturbed member (CF, continuous line), ensemble mean (EM, dash) and perturbed members 1,10,20,30,40,50 (dot-dash lines). Left panel refers to southern extra-tropics, middle panel to tropics, right panel to the northern extra-tropics. Bottom row: as top row for the forecast evolution of the standard deviation of forecast anomaly (forecast activity, SDAF). Statistics accumulated over the 2023-10-01 to 2024-01-20 period. Verification against the ECMWF operational analysis.

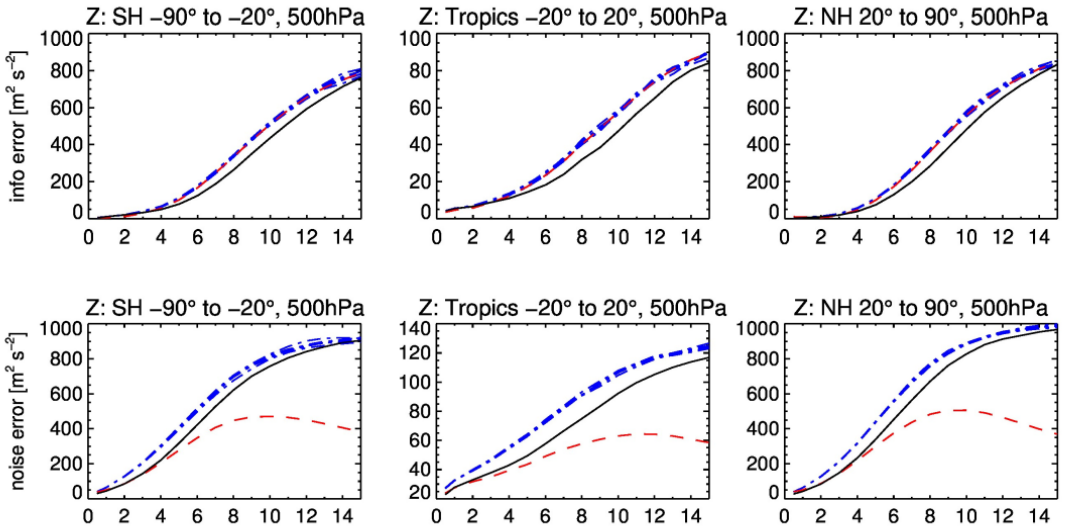

Figure 5: Top row: Evolution of the ENS Prediction System Information Error (IE) of 500 hPa geopotential (Z) for the unperturbed member (CF, continuous line), ensemble mean (EM, dash) and perturbed members 1,10,20,30,40,50 (dot-dash lines). Left panel refers to southern Z500 forecast fields for the CF, EM and perturbed forecasts. Bottom row: as top row for the forecast evolution of Noise Error (NE). Statistics accumulated over the 2023-10-01 to 2024-01-20 period. Verification against the ECMWF operational analysis.

4. Information and Noise Error for NWP development

The application of the Information and Noise Error decomposition can be potentially useful in the research and development workflow of NWP as it provides additional insight into the added value (or lack thereof) of proposed changes to the analysis and forecast system. We illustrate this by presenting some relevant examples from recent development work at ECMWF.

4.1. Revised vertical diffusion scheme in the stratosphere

This change to the IFS model involved reducing the vertical diffusion in stable conditions above the tropopause. The motivation of this change stems from the fact that the gradient functions used for diffusion of momentum, heat and tracers, such as moisture, are empirically derived from near surface observations. Due to a lack of similar observations or analyses at higher altitudes, the same functions are used throughout the entire depth of the atmosphere in stable conditions, including across the tropopause and into the stratosphere. Gradients in winds and temperatures have been shown to be too weak across the tropopause, indicative of excessive diffusion into the stable stratosphere. The unphysically strong vertical diffusion is known to have deleterious impact on the modelled quasi-biennial oscillation (QBO) of the tropical winds in stratosphere, reducing not only the vertical wind shear but also the QBO amplitude, thus increasing the mean wind biases in the tropical stratosphere (Polichtchouk et al., 2021; section 2.5).

The verification of the proposed change in the ECMWF data assimilation cycle showed generally positive impacts with improved fit of the 4D-Var analysis and first guess forecasts from the background state to most observation types, particularly visible in the upper troposphere and lower stratosphere (UTLS; not shown). Forecast medium-range impact showed generally positive results in the UTLS in the tropics but mixed results in the extra-tropics (both against operational ECMWF analysis and radiosonde observations) for both temperature (not shown) and wind components (Figure 6, left panel). On the other hand, the Information Error plots (Figure 7, left panel) show that the degradation in the extra-tropical UTLS mostly disappears in this metric. The increase in RMSE comes exclusively from the increase in the Noise Error component (Figure 7, right panel) and it is largely driven by the increase in forecast activity of the modified IFS (Figure 6, right panel). which indicates a forecast with a more realistic level of variability.

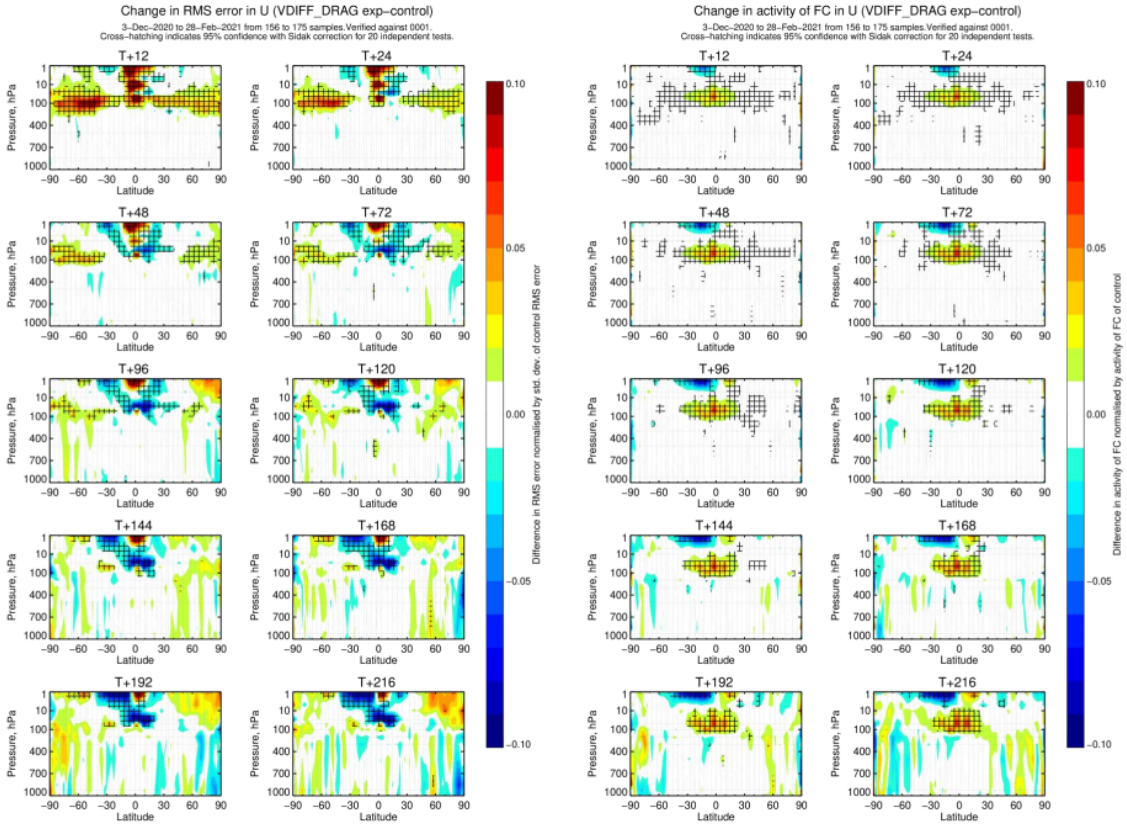

Figure 6: Normalised difference in the RMSE (left panel) and forecast activity (SDAF) of the zonal wind field forecast of the experiment with the reduced vertical diffusion vs a control. Verification versus ECMWF operational analysis over the 3/12/2020 to 28/02/2021 period. Negative values indicate reduced RMSE (SDAF) in the experiment with reduced vertical diffusion.

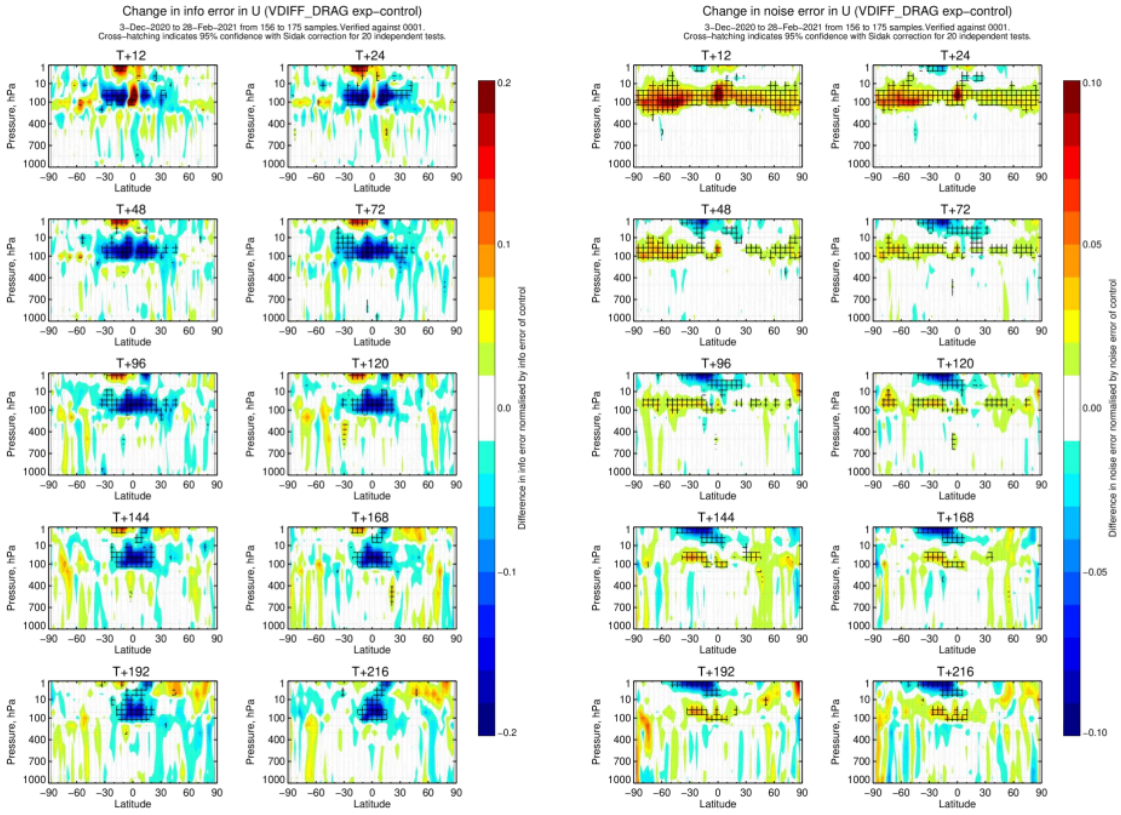

Figure 7: Same as Figure 6 for the Information Error (IE, left panel) and Noise Error (NE) measures.

4.2. Off-centring of the IFS Semi-implicit time-stepping scheme

The spectral, semi-implicit, semi-Lagrangian dynamical core of the IFS, described in ECMWF (2024), employs a fully centred second order time-stepping scheme. The IFS formulation and code allow the option to off-centre the semi-implicit time weights from their default value 1/2, 1/2 to (1-ε)/2, (1+ε)/2 where the first weight is applied on the explicit and the second on the implicit part of the time discretization respectively. A small amount of off-centring has been introduced here, equal to 0.05, to control the resonant growth of spurious modes generated by sharp gradients in flows whose Courant number exceeds unity, and which arise especially at high horizontal resolution in the IFS in the stratosphere if the vertical resolution is also not concomitantly increased. Such spurious modes can lead to mean biases as discussed in Polichtchouk et al. (2019).

In terms of impact of the assimilation system, the Off-Centring experiment shows significantly improved analysis and first guess from background fits to all stratospheric sounding observations (not shown). This is a clear indication that this model change is beneficial in the data assimilation cycle in extracting more information from the current observing system.

In terms of impact on forecast performance, standard metrics like ACC (not shown) and RMSE show significant reductions in the stratosphere, e.g., Figure 8, left panels. However, the reduction in forecast RMSE appears associated with collocated reductions in forecast activity (Figure 8, right panel). The plots showing the evolution of Information and Noise Error components (Figure 9) indicate that while the Off-Centring is successful in reducing the Noise component of the error budget, there are degradations in terms of Information Error, e.g. in the south hemisphere stratosphere, that merit further investigation. Interestingly the IE component shows localised degradations in the tropospheric mid-latitudes in the early forecast range, indicating that while this model change is targeting stratospheric effects, its impacts are broader.

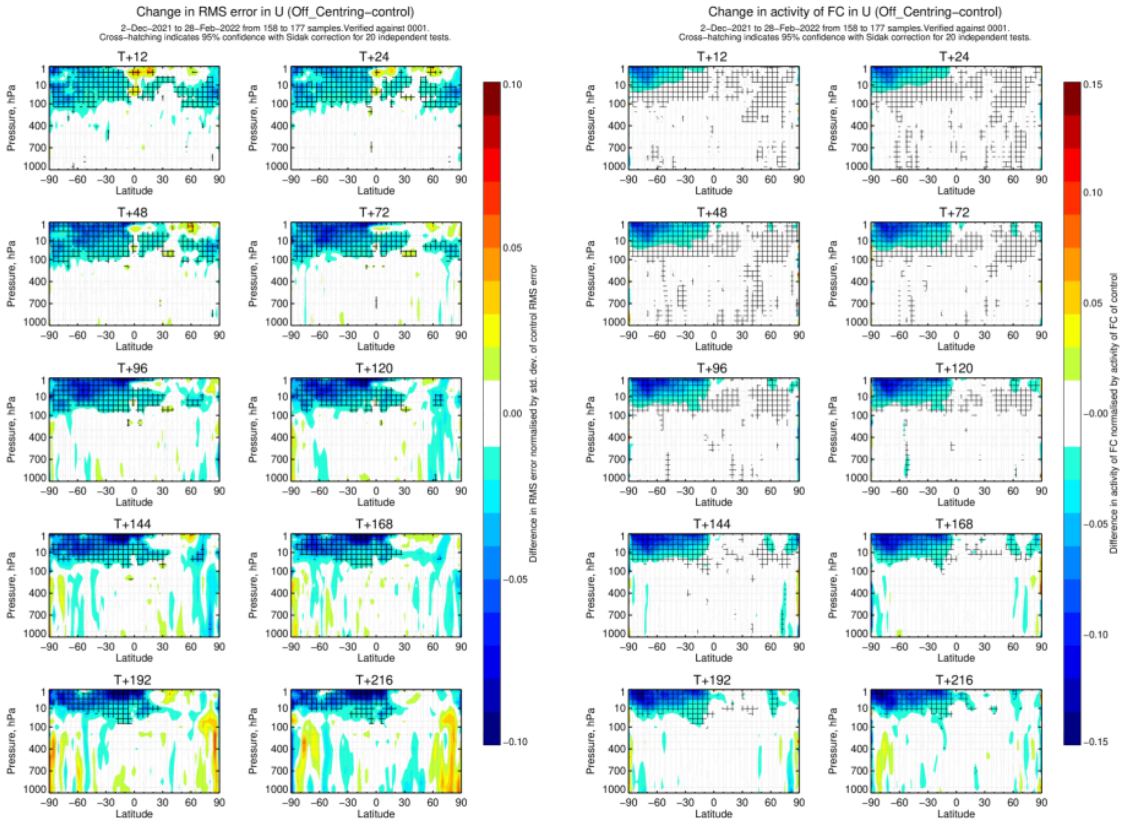

Figure 8: Normalised difference in the RMSE (left panel) and forecast activity (SDAF, right panel) of the zonal wind forecast of the Off-Centring experiment vs a control. Verification versus ECMWF operational analysis over the 2/12/2021 to 28/02/2022 period. Negative values indicate reduced RMSE (SDAF) in the Off-Centring experiment.

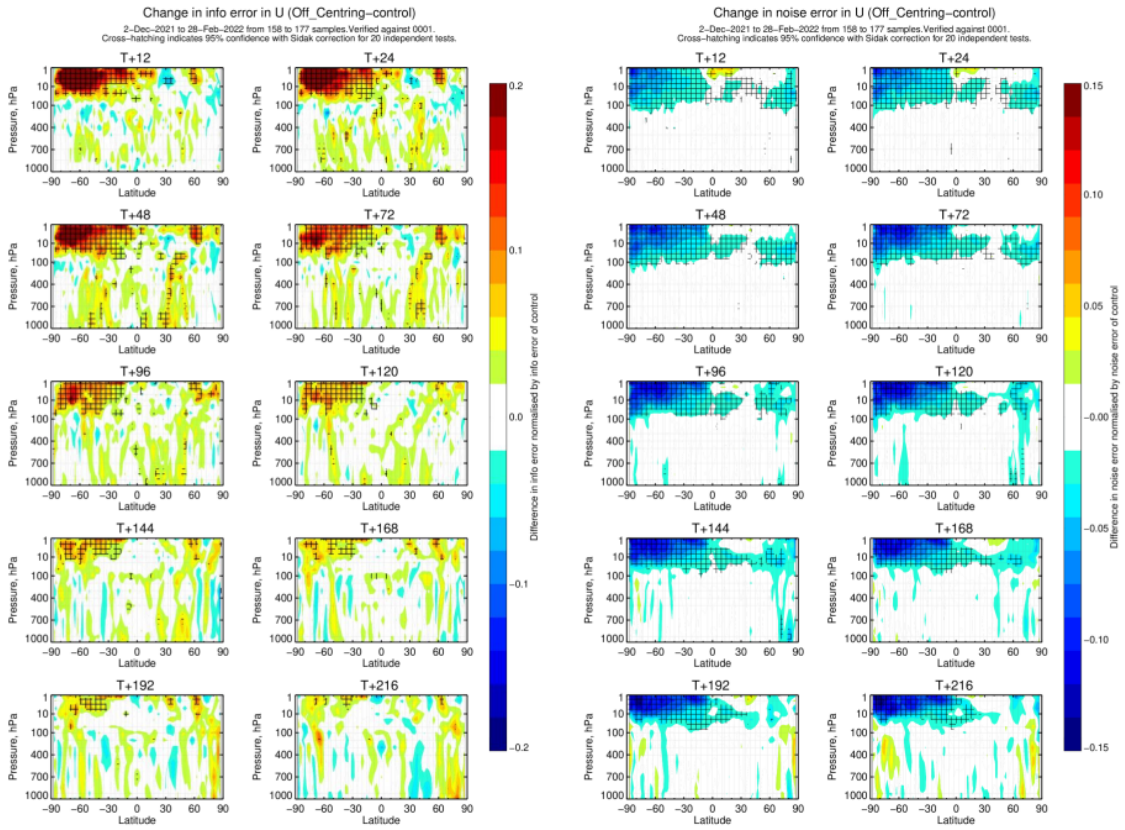

Figure 9: Same as Figure 7 for the Information Error (IE, left panel) and Noise Error (NE, right panel) measures.

4.3. Impact of ATMS observations

A standard activity in operational NWP development is the introduction of new observing systems that become available for use in an operational data assimilation context, or the improved exploitation of already available and used observing systems. As an example of this type of NWP activities we consider here the forecast impact of adding observations from the Advanced Technology Microwave Sounder (ATMS) instrument on board various polar orbiters (JPSS, NOAA- 20, NOAA-21) on a baseline system which ingests all other currently ingested observations. Over the test period (Dec 2022 – Feb 2023) the experiment which adds the ATMS radiances shows generalised improvements in the first guess departures (O-B) of the order 0.5% against most of the other observing systems (not shown). These improvements are reflected in improvements in standard verification measures like RMSE (Figure 10, left panel), most visible in the stratosphere and in the south hemisphere troposphere. The forecast activity (SDAF, Figure 10, right panel) is largely unchanged. The Information Error and Noise Error plots (Figure 11) show that both components of the error budget contribute to the reduction in RMSE, indicating that the addition of ATMS observations to the global observing system does indeed improve the statistical resolution of the forecast system and at the same time reduces the noise error component.

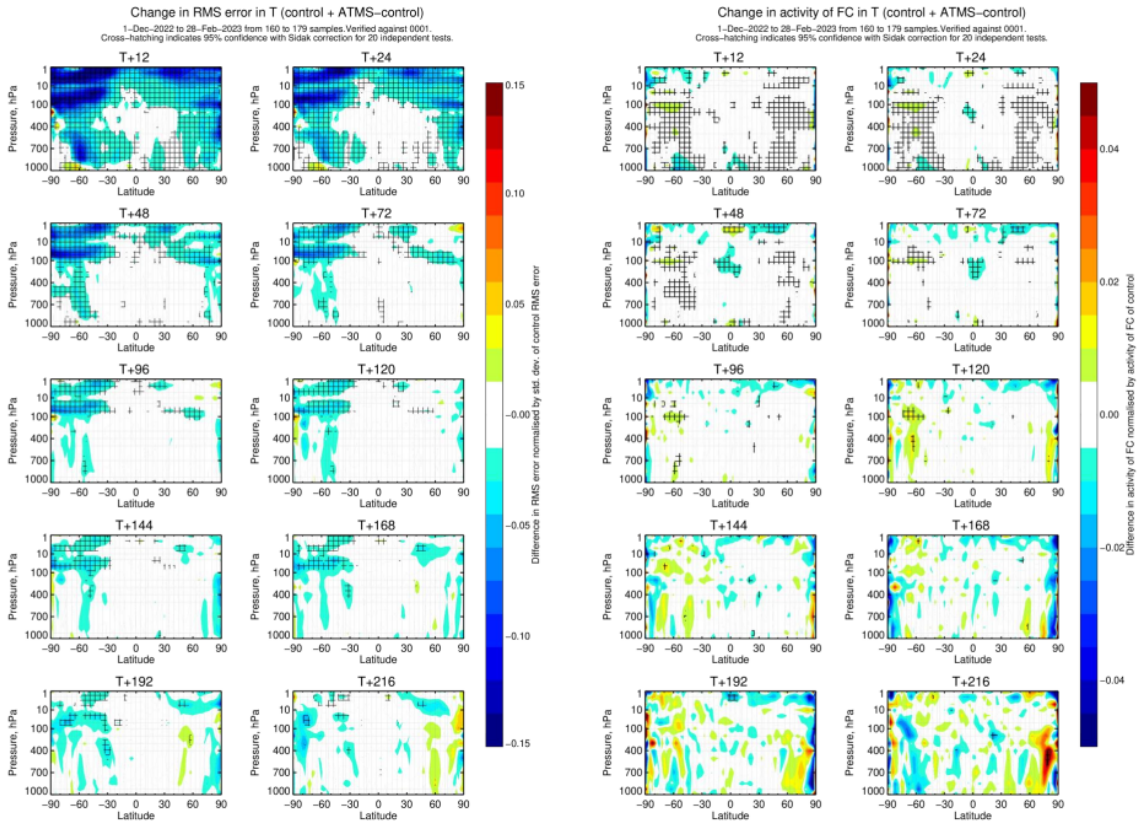

Figure 10: Normalised difference in RMSE (left panel) and forecast activity (SDAF, right panel) of the temperature forecast of the ATMS experiment vs its control. Negative values indicate reduced RMSE (SDAF) for the ATMS experiment. Verification against ECMWF operational analysis over the 1/12/2022 to 28/02/2023 period.

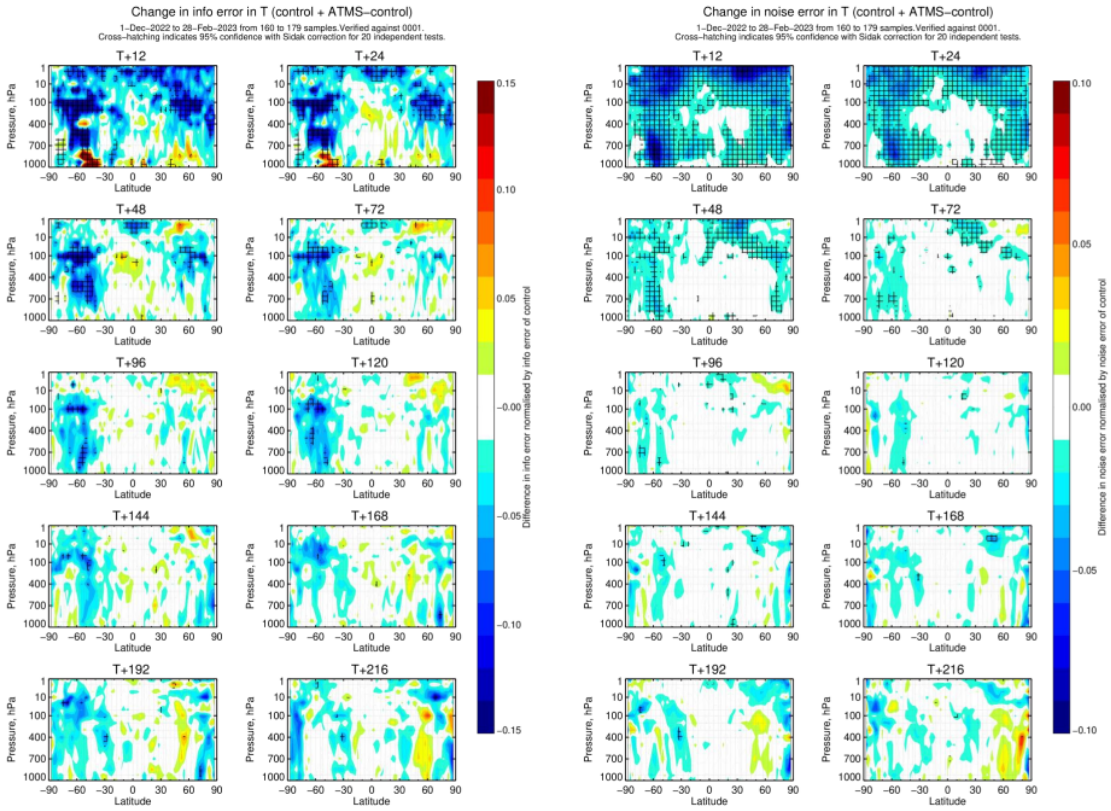

Figure 11: Same as Figure 9 for the Information Error (IE, left panel) and Noise Error (NE, right panel).

4.4. Machine Learning Forecast Models

A well-known feature of the first generation of Machine Learning "deterministic" (one-shot) weather emulators (e.g., GraphCast, Lam et al., 2022; FourCastNet, Pathak et al., 2022; Pangu-Weather, Bi et al., 2023; etc.) is that they produce overly smooth and unphysical forecasts with reduced spectral energy and also reduced forecast activity at increasing lead times (Bonavita, 2024). While this is often mentioned as a possible caveat in the interpretation of their forecast performance measures, it has so far proved difficult to quantify its impact. The application of the IE/NE decomposition can shed some light on this issue. As an example, we show in Figure 12 the forecast RMSE (top row), Forecast Activity (SDAF, second row), Noise Error (NE, third row) and Information Error (bottom row) of 500 hPa geopotential for GraphCast (black lines) and ECMWF operational forecasts (red lines) over the March 2024 – February 2025 period. In a traditional type of forecast skill measure like RMSE, GraphCast appears to provide superior performance over the ECMWF operational forecasts in all areas. When a more precise measure of statistical resolution such as IE is used, the picture is mixed. In the troposphere, where forecast activity of Graphcast is at realistic levels, even though decreasing with forecast lead time, the performance advantage of GraphCast (around 12 hours in added predictive skill) is confirmed by the IE measure.

On the other hand, when one looks at surface variables (e.g., T2m in Figure 13) the IE/NE decomposition indicates that the ECMWF operational forecasts have comparable or better statistical resolution, and thus intrinsic forecast skill, than those produced by the GraphCast ML emulator. This conclusion extends to other surface variables (e.g., 10 meter u/v winds, Mean Sea Level Pressure) and indicates that for these variables a significant contributor to the advantage of GraphCast in terms of RMSE (and ACC) derives from its relative reduction in forecast activity with respect to the ECMWF IFS.

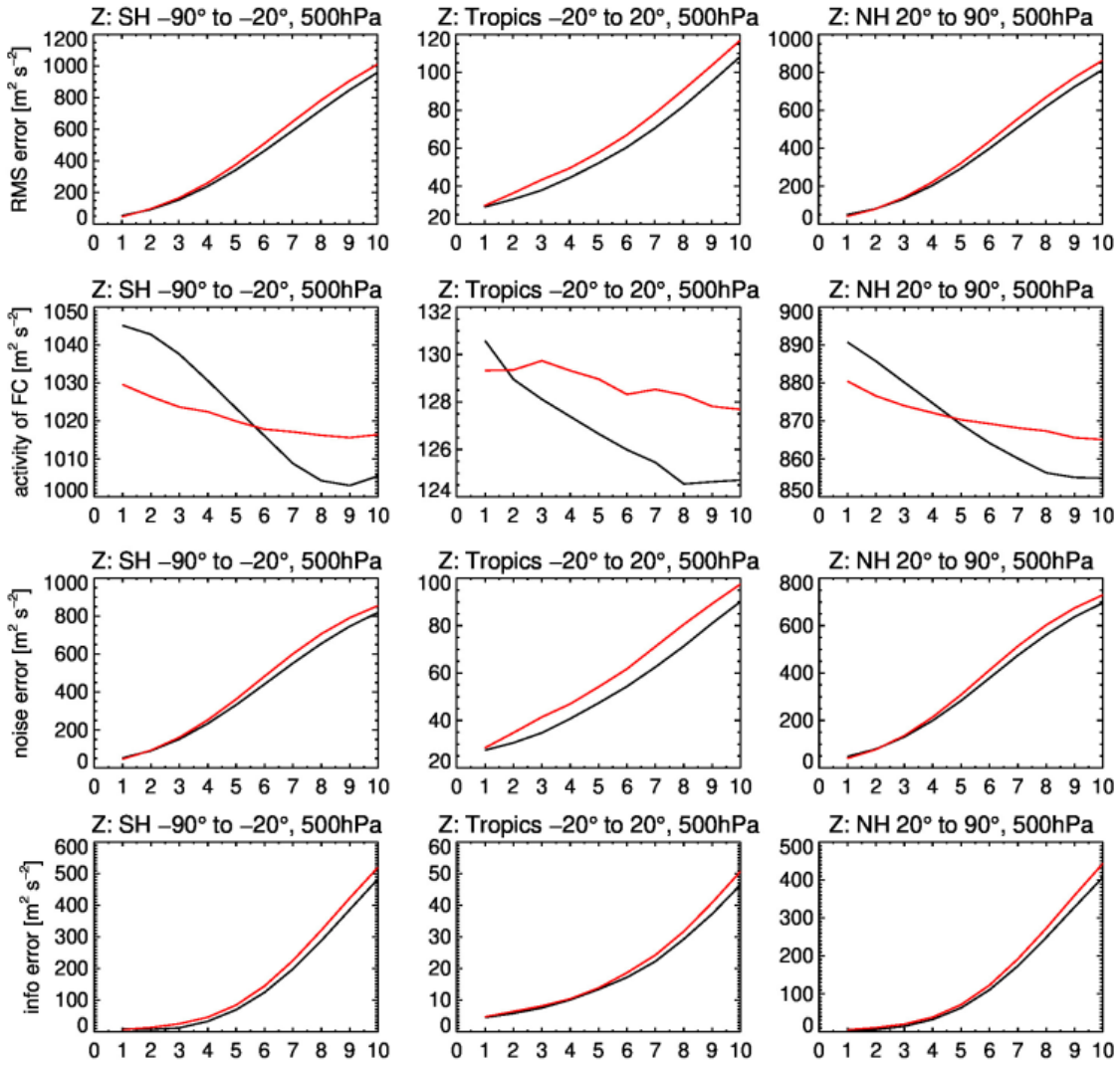

Figure 13: Geopotential 500hPa forecast RMSE (first row), forecast activity (second row), Noise Error (NE, third row) and Information Error (IE, bottom row) for GraphCast (black lines) and ECMWF operational forecasts (red lines) over the southern extra tropics (left column), tropics (middle column) and northern extra tropics (right column). Verification versus ECMWF operational analysis over the 1/03/2024 to 28/02/2025 period.

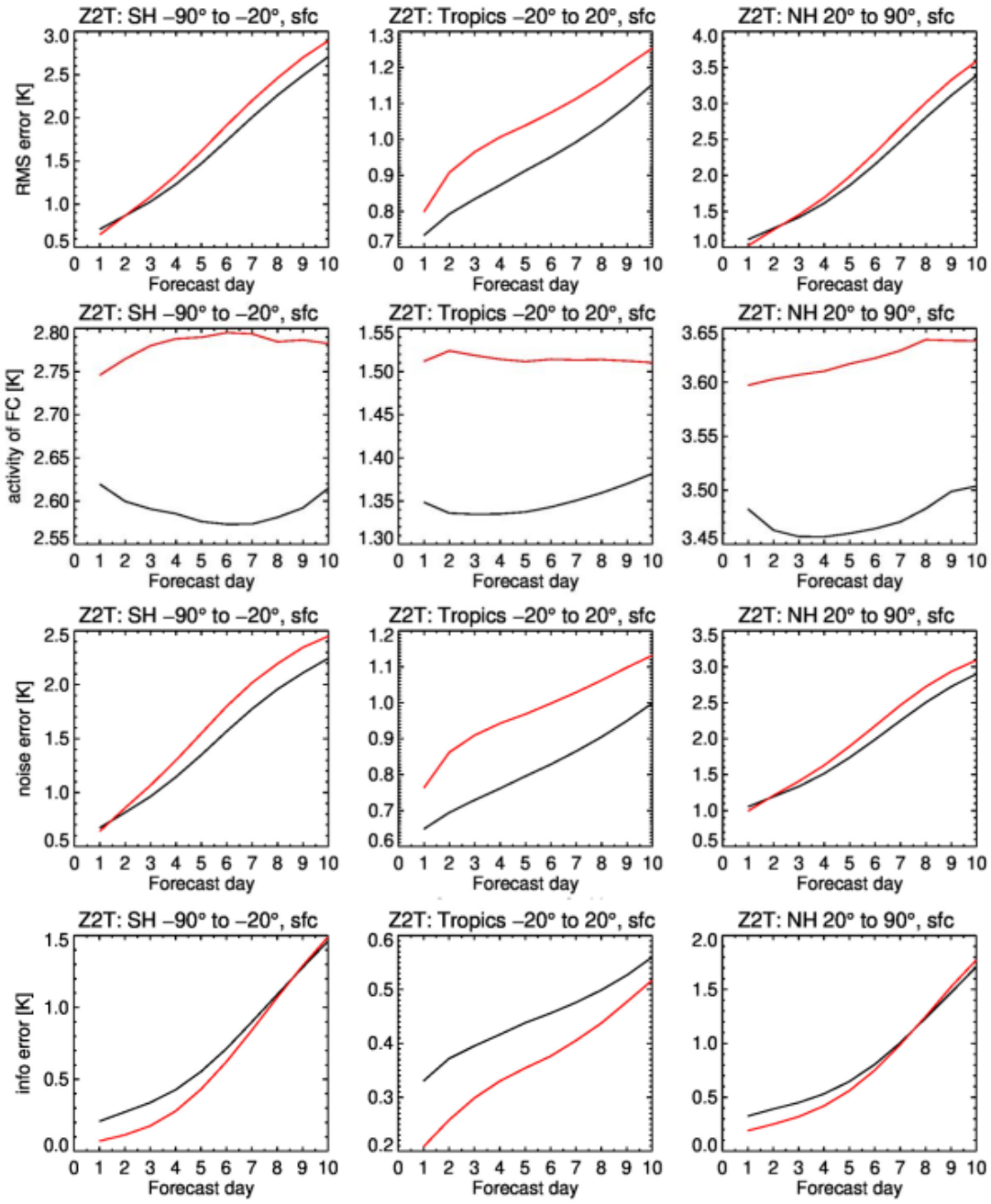

Figure 14: As in Figure 13 for 2 metre temperature (T2m) forecasts.

5. Conclusions

In this note we have started from the examination of two recently proposed forecast skill measures, the Forecast Information and Forecast Noise (FTZP24), which have been devised to give an undiluted estimate of the statistical resolution of a forecast system. With respect to the original formulation of FI and FN in FTZP24, a few modifications have been introduced: a) we are using debiased anomalies in their definition; b) we are using the signed square root of the FI and FN defined in FTZP24; and c) we have extended the framework to allow for area-weighting in the computation of these forecast scores. With these modifications, FI and FN can be expressed as functions of standard forecast measures (ACC, SDAF, SDAV) and effectively synthetise information present in these quantities. Note that if the anomalies are not debiased, as in FTZP24, then the equivalent decomposition applies to the RMSE and the non-debiased versions of the activity and anomaly correlation but is equivalent. Specifically, FI can be viewed as a rescaled form of ACC (i.e., Eq. 11) where the covariance between forecast and observed anomaly is normalised by the variance of the observed anomaly only. This choice has the advantage of penalising overly smooth forecasts which are rewarded by the well-known "double penalty" effect. On the other hand, FI cannot be guaranteed to have an upper bound of one as ACC, and this fact can make its interpretation as a forecast skill measure difficult. For this reason, we have developed two related error measures, Information Error and Noise Error, that avoid the pitfalls described above and provide a meaningful and unambiguous decomposition of the forecast STDE for debiased anomalies (and RMSE for undebiased anomalies). Specifically, the IE/NE decomposition projects the total forecast error into components parallel and orthogonal to the observed anomaly. While both components need to be minimised to reduce forecast error, the reduction of the IE is more crucial and possibly more difficult to achieve. In particular, damping forecast activity will increase IE even when it reduces the overall RMSE. In this sense, one can interpret IE as a measure of the statistical resolution of the forecast system, i.e. the ability of the forecast system to distinguish outcomes conditioned on forecast outputs (Wilks, 2011; TTZ05).

The rest of this note is devoted to applications of these new measures to better understand their characteristics and potential use cases. We first look at applying FI/FN to ECMWF operational medium-range forecast system. Results are consistent with results shown in FTZP24 and with the general understanding of current predictive skill of the NWP operational ensemble prediction systems. An interesting result of this exercise has been that simple post-processed products like the ensemble forecast mean have the same statistical resolution (i.e., information content) of the individual perturbed forecasts and lower statistical resolution than the unperturbed forecast. These results further support, in both quantitative and theoretical aspects, the conclusion that the advantage of ensemble mean/median products in standard forecast metrics (ACC, RMSE) with respect to the unperturbed forecast is based on their reduced forecast activity. On the other hand, our results also indicate that ensemble forecast mean has the lowest STDE for a given ACC (Figure 3), which supports the widely

held assumption that ensemble averaging is an effective tool of removing noise from scales that become progressively less predictable for the forecast system in use.

Another area of application of the new measures is in the standard NWP development cycle. For illustration, we present a few cases of recent upgrades proposed for the ECMWF Integrated Forecast System (IFS) model. Both upgrades introduce non negligible changes in the IFS forecast activity and the use of IE/NE is shown to allow a more quantitative evaluation of the improvement/degradation of the IFS statistical resolution as a result of these changes than what is possible with standard forecast skill measures. This is also true for the evaluation of the intrinsic forecast skill of Machine Learning Weather Prediction models, where the impact of less active, more diffusive forecasts on the statistical resolution of the forecast system needs to be assessed in a quantitative manner.

The forecast error metrics described here are new additions to the set of NWP verification tools and more work is needed to better understand their capabilities and limitations. For example, the verification (truth) used in this note is model based (ECMWF operational analysis and/or own analysis verification). There are examples where this approach can lead to misleading results, e.g. when estimating changes due to changes in the observing system, which will also lead to changes in the activity of the verifying analysis. For these cases, it would be preferable to use independent observations. This and other refinements are left for further development.

Finally, we emphasize that all the metrics described here should be used in combination with other forecast evaluation metrics to make an informed decision about proposed changes to the forecasting system. For example, both IE and NE are sensitive to changes in conditional/unconditional systematic errors whose reduction is also an important target of forecast system development. For the end user of NWP forecast products, the accuracy of the products in terms of minimising and quantifying their expected error will always be of paramount importance. For NWP and Machine Learning model developers, however, the ability to distinguish between the resolution and reliability components of the error budget will be crucial in order to understand whether new systems deliver improvements in the information content of the forecasts or a still useful, but less fundamental, improvement in their reliability. This especially in light of the fact that while in the NWP workflow calibration of raw model output is typically done in a separate post-processing step accessing a training dataset of forecast/verification pairs, in current ML model development the calibration step is absorbed in the model development phase and the loss functions more commonly used aim to minimise total error measures (e.g., RMSE, CRPS).

Acknowledgments

This work has greatly benefited from discussions and feedback from Zoltan Toth (Global Systems Laboratory, NOAA, Boulder, USA), Jie Feng (Fudan University, Shanghai, China) and Inna Polichtchouk, Michail Diamantakis, Annelize Van Niekerk, Peter Dueben, Martin Leutbecher and Thomas Haiden (ECMWF, Reading, UK).

Appendix A

Compared to the definition of information and noise in FTZP24, and as presented in the introduction here, the practical computation of these statistics makes two changes for consistency with the way the decomposition of forecast error into activity and anomaly correlation is normally computed at weather forecasting centres:

1. The mean of the anomaly is subtracted (equivalently debiased)
2. Area weighting is performed.

Here we first present the statistics in a more standard summation form, representing exactly how the calculations are done in the code of the Iver forecast verification tool at ECMWF. Then we show how these modifications can be understood geometrically and how these statistics parallel the derivations given in the main text for the simpler case (the decomposition of RMSE without area weighting).

Statistics are first computed on the level of a single forecast valid at a specific time, on one pressure level, and over a predefined region such as the northern hemisphere. These single-forecast statistics are later aggregated using the unweighted mean (see below). This allows the formation of populations for statistical significance testing as described by Geer (2016). For a single forecast field and region as described, the basis of computation is the bias-corrected anomaly for forecast $d_{f,i}$ and for the reference or 'truth' analysis $d_{t,i}$, at one location i out of the n locations composing the field:

$$d_{f,i} = x_{f,i} - x_{c,i} - \sum_{i=1}^{n} w_i \, (x_{f,i} - x_{c,i}) \qquad \text{(A1)}$$

$$d_{t,i} = x_{t,i} - x_{c,i} - \sum_{i=1}^{n} w_i \, (x_{t,i} - x_{c,i}) \qquad \text{(A2)}$$

The summation is the grid box area-weighted regional mean anomaly where the weights are proportional to the grid box area, which is computed using grid box central latitude λ_i and the latitude width $\Delta\lambda_i$ as follows, with a a normalising constant, such that:

$$w_i = a \sin \Delta \lambda_i \cos \lambda_i \qquad \text{(A3)}$$

$$\sum_{i=1}^{n} w_i = 1 \qquad \text{(A4)}$$

The forecast and truth activity are here the area-weighted sample standard deviations of the relevant anomaly, calculated as

$$\mathrm{SDAF} = \sqrt{\sum_{i=1}^{n} w_i \left(d_{f,i}\right)^2} \qquad \text{(A5)}$$

$$\mathrm{SDAV} = \sqrt{\sum_{i=1}^{n} w_i \left(d_{t,i}\right)^2} \qquad \text{(A6)}$$

The area-weighted dot product of the forecast and truth anomalies, p, is common to the anomaly correlation and the information:

$$p = \sum_{i=1}^{n} w_i \, d_{f,i} d_{t,i} \qquad \text{(A7)}$$

So that

$$\mathrm{ACC} = \frac{p}{\mathrm{SDAF} \times \mathrm{SDAV}} \qquad \text{(A8)}$$

$$\mathrm{FI} = \frac{p}{\mathrm{SDAV}^2} \qquad \text{(A9)}$$

$$\mathrm{IE} = abs(1 - \mathrm{FI}) \times \mathrm{SDAV} \qquad \text{(A10)}$$

And to compute the length of the noise vector, starting from a simple vector calculation from Figure 1c, but applying area weighting:

$$r = \sum_{i=1}^{n} w_i \left(d_{f,i} - \mathrm{FI}\, d_{t,i}\right)^2 \qquad \text{(A11)}$$

so that noise error NE is computed as

$$NE = \sqrt{r} \qquad \text{(A12)}$$

As mentioned, the measures presented in Iver and in the figures in this paper are then aggregated using a simple mean over the sample of forecasts generated from m different base times j so that we are usually showing the sample mean information, noise, anomaly correlation and activities as follows, based on the information and noise calculations for each individual forecast j, where the index j has been omitted so far for clarity:

$$\overline{FI} = \frac{1}{m} \sum_{j=1}^{m} FI_j \qquad \text{(A13)}$$

$$\overline{\mathrm{NE}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{NE_j} \qquad \text{(A14)}$$

$$\overline{\mathrm{IE}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{IE_j} \qquad \text{(A15)}$$

$$\overline{\mathrm{ACC}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{ACC_j} \qquad \text{(A15)}$$

$$\overline{\mathrm{SDAF}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{SDAF_j} \qquad \text{(A16)}$$

$$\overline{\mathrm{SDAV}} = \frac{1}{m}\sum_{j=1}^{m} \mathrm{SDAV_j} \qquad \text{(A17)}$$

The figures we show are all based on these scores that have been meaned over forecasts (equivalently cycles). This averaging breaks the exact mathematical decomposition of RMSE or STDE into the other scores that we have presented, and which is strictly valid for a single cycle and forecast only. However, if the statistics do not have large variations with time, it can be shown mathematically that the errors are small (this is not included here as it would require an additional long appendix). We also see that in practice, and especially for scores like Z500, the decomposition remains reasonably closely valid, even when applied to the time-mean scores.

It is not immediately obvious how the area-weighted, mean-subtracted versions of these scores follow the geometric interpretation presented in the introduction. To make the link, we first need to think about how removing a mean would be represented in the geometric approach. To do this, we can seek a decomposition of a vector **a** into two vectors **a′** and **m**

$$\mathbf{a} := \mathbf{a}' + \mathbf{m} \qquad \text{(A18)}$$

We then impose two conditions: that **a′** and **m** should be orthogonal, and that the $\mathbf{m} = (m, m, m \ldots\ldots m)$ is the same length m in every dimension. This scalar m is our desired mean.

To include area weighting, we also need to generalise the dot product, which we will denote

$$< \mathbf{a}, \mathbf{b} > = \sum_{i=1}^{n} w_i a_i b_i \qquad \text{(A19)}$$

where a_i and b_i are the distances along each unit vector. This notation reflects that in a broader sense we are defining an alternative inner product with which to measure length.

To impose orthogonality of the mean in the case of a weighted inner product, we first take the inner product of **a′** with itself, relying on this product being distributive, which is easily shown for the above specific definition. Hence, we obtain:

$$< \mathbf{a}', \mathbf{a}' >=< (\mathbf{a}-\mathbf{m}), (\mathbf{a}-\mathbf{m}) >=< \mathbf{a}, \mathbf{a} > -< \mathbf{2m}, \mathbf{a} > +< \mathbf{m}, \mathbf{m} > \quad \text{(A20)}$$

If $\mathbf{a}'$ and $\mathbf{m}$ were orthogonal according to this weighting, then Pythagoras' rule would hold, i.e.

$$<\mathbf{a}', \mathbf{a}' >=< \mathbf{a}, \mathbf{a} > -< \mathbf{m}, \mathbf{m} > \quad \text{(A21)}$$

It is therefore possible to achieve this orthogonality between $\mathbf{a}'$ and $\mathbf{m}$ if

$$< \mathbf{m}, \mathbf{a} >=< \mathbf{m}, \mathbf{m} > \quad \text{(A22)}$$

Applying the weighted dot product (A19) to a debiased vector (A20) therefore **defines** the mean as follows

$$\sum_{i=1}^{n} w_i m a_i = \sum_{i=1}^{n} w_i m^2 \quad \text{(A23)}$$

So that

$$m = \frac{\sum_{i=1}^{n} w_i a_i}{\sum_{i=1}^{n} w_i}. \quad \text{(A24)}$$

Since we have also defined that our weights sum to 1, we can ignore the denominator. We have done no more than recover the usual definition of a weighted mean, but we can now show how and why it is the appropriate definition of the mean to be used with the weighted distance measure represented by our generalised dot product (inner product).

A consequence of defining the weighted inner product is that it implies that if we just rescale our vectors appropriately, we can use the standard Cartesian dot product and other geometrical tools. If the original forecast and truth fields are denoted x_i as in the main text, then the weighted mean that we derived above only makes sense if our vectors are not rescaled but simply defined $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$. (unlike the main text, here we do not use " to denote this). However, there is another set of vectors where

$$\mathbf{a}^* = (\sqrt{w_1}a_1, \sqrt{w_2}a_2, \ldots \sqrt{w_n}a_n). \quad \text{(A25)}$$

and we are applying a scaling transformation to the vectors. With this scaling, the standard dot product is the weighted inner product that we applied in the unscaled vectors:

$$\mathbf{a}^* \cdot \mathbf{b}^* = \sum_{i=1}^{n} a_i^* b_i^* = \sum_{i=1}^{n} w_i a_i b_i = <\mathbf{a}, \mathbf{b}>. \qquad \text{(A26)}$$

(This broadly links to the notion of an inner product space, i.e. a space defined by the inner product, but it is much easier to think about scaling within one space in this case). With these scaled vectors we can apply the standard dot product and cosine triangle rule. We have already effectively made use of this alternative scaling (or alternative space) in the main text when, in order to make the RMSE equal to the length according to the dot product, we defined our vectors from x_i to include a constant scaling $w_i = \frac{1}{\sqrt{n}}$ for all i.

To apply these concepts to forecast verification, we define the following scalars representing a weighted mean of each vector in the triangle formed by the forecast, truth and climatology:

$$m_{fc} = \sum_{i=1}^{n} w_i \, (x_{f,i} - x_{c,i}) \quad \text{(A27)}$$

$$m_{tc} = \sum_{i=1}^{n} w_i \, (x_{t,i} - x_{c,i}) \quad \text{(A28)}$$

$$m_{ft} = \sum_{i=1}^{n} w_i \, (x_{f,i} - x_{t,i}) \quad \text{(A29)}$$

We can then define vectors to represent the bias corrected forecast anomaly, true anomaly and forecast error as follows:

$$\mathbf{d}_f = \boldsymbol{x_f} - \boldsymbol{x_c} - \mathbf{m}_{fc} \qquad \text{(A30)}$$

$$\mathbf{d}_t = \boldsymbol{x_t} - \boldsymbol{x_c} - \mathbf{m}_{tc} \qquad \text{(A31)}$$

$$\mathbf{d}_{ft} = \boldsymbol{x_f} - \boldsymbol{x_t} - \mathbf{m}_{ft} \qquad \text{(A32)}$$

It is clear that these three new vectors also form a triangle, because

$$\mathbf{d}_f - \mathbf{d}_t = \boldsymbol{x_f} - \boldsymbol{x_t} - \left(\mathbf{m}_{fc} - \mathbf{m}_{tc}\right) = \boldsymbol{x_f} - \boldsymbol{x_t} - \mathbf{m}_{ft} \qquad \text{(A33)}$$

so that

$$\mathbf{d}_{ft} = \mathbf{d}_f - \mathbf{d}_t \qquad \text{(A34)}$$

This is illustrated in Figure A1 for two dimensions.

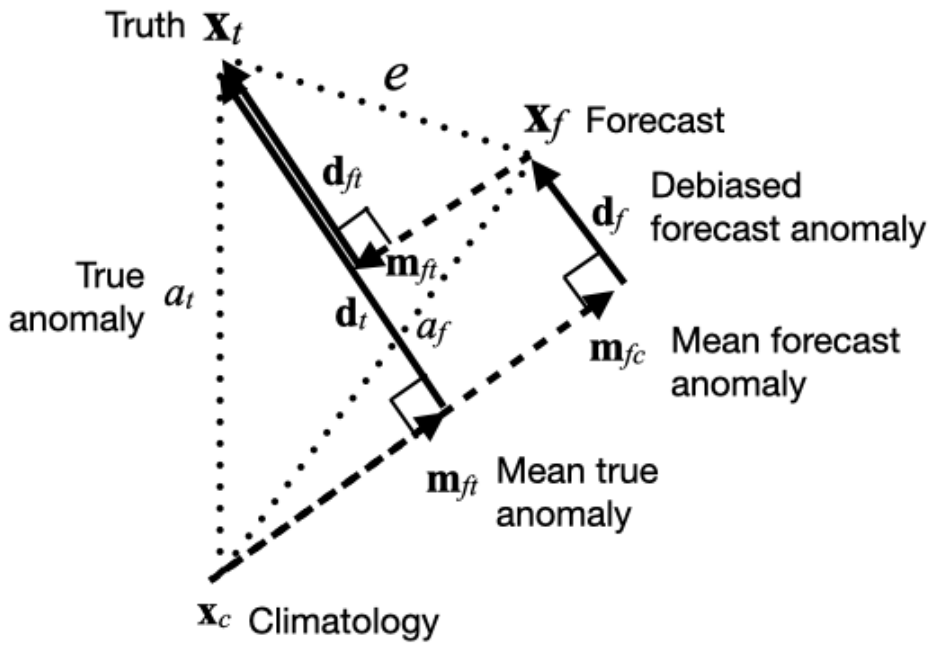

Figure A1: Visual representation of the relationship between forecast and observed mean and debiased anomalies. See text for details.

Note that the mean vectors are always just scalar multiples of the diagonal vector (1,1,1,..1) in however many dimensions. It is clear from the figure that in 2 dimensions, $\mathbf{d}_{ft} = \mathbf{d}_f - \mathbf{d}_t$ makes a degenerate triangle. In other words, for a two-dimensional space, debiasing the anomalies makes the angle between the anomalies equal to 0 and the anomaly correlation equal to one. This is linked to the Bessel correction for computing standard deviation (the use of $n - 1$ rather than n), which recognises there is one less degree of freedom available once the mean has been subtracted. Note that no statistics computed here include the Bessel correction, since it is assumed that in practice n is large enough that $n \approx n - 1$.

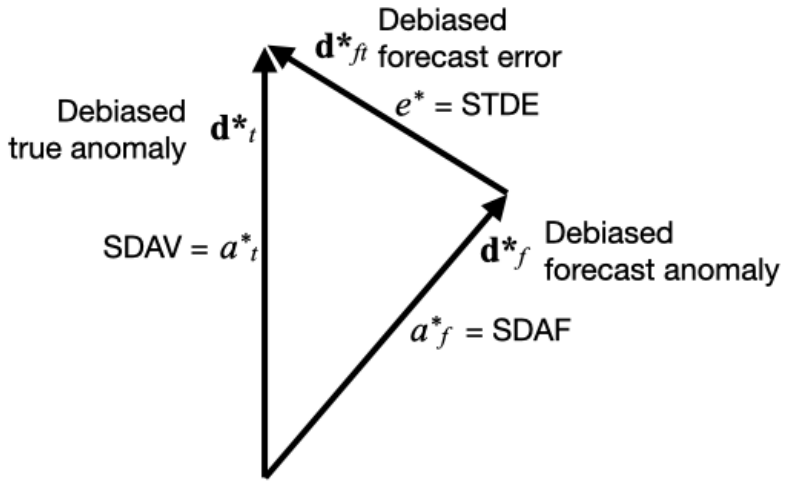

Figure A2: Visual representation of the relationship between debiased anomalies and error. See text for further details.

Figure A2 then exemplifies the triangle formed by the debiased, rescaled anomalies (vectors denoted by *) to take account of the area weighting. This is now notionally for a three dimensional (three point) forecast field where the triangle of Figure A1 is no longer degenerate and we can rotate our viewpoint to again put the debiased, rescaled true anomaly along the y-axis of the figure. The length of the forecast error vector will be denoted e^*, and similarly a_t^* and a_f^* denote the lengths of the true and forecast anomaly after debiasing and rescaling (area weighting). By expanding out these lengths in summation form we can show that they correspond to the standard deviation of forecast error (STDE), the standard deviation of the true and forecast anomaly (SDAV and SDAF) and that the angle between the latter two is the anomaly correlation coefficient ACC:

$$e^* = \sqrt{\mathbf{d}^*_{\mathbf{ft}} \cdot \mathbf{d}^*_{\mathbf{ft}}} = \sqrt{\sum_{i=0}^{n} w_i \left(x_{f,i} - x_{t,i} - m_{ft}\right)} = \text{STDE} \qquad \text{(A35)}$$

$$a_t^* = \sqrt{\mathbf{d}^*_{\mathbf{t}} \cdot \mathbf{d}^*_{\mathbf{t}}} = \sqrt{\sum_{i=0}^{n} w_i \left(x_{t,i} - x_{c,i} - m_{tc}\right)} = \text{SDAV} \qquad \text{(A36)}$$

$$a_f^* = \sqrt{\mathbf{d}^*_{\mathbf{f}} \cdot \mathbf{d}^*_{\mathbf{f}}} = \sqrt{\sum_{i=0}^{n} w_i \left(x_{f,i} - x_{c,i} - m_{fc}\right)} = \text{SDAF} \qquad \text{(A37)}$$

The triangle cosine rule applied to the triangle $\mathbf{d}^*_{\mathbf{ft}} = \mathbf{d}^*_{\mathbf{f}} - \mathbf{d}^*_{\mathbf{t}}$ defines the decomposition of the STDE into SDAV, SDAF and ACC as follows, with ϕ being the angle between the true and forecast anomaly, after debiasing and rescaling.

$$(e^*)^2 = (a_t^*)^2 + \left(a_f^*\right)^2 - 2a_t^* a_f^* \cos\phi \qquad \text{(A38)}$$

or equivalently

$$(\mathrm{STDE})^2 = (\mathrm{SDAV})^2 + (\mathrm{SDAF})^2 - 2\mathrm{SDAV}\times\mathrm{SDAF}\cos\phi \qquad \text{(A39)}$$

The cosine dot product rule hence defines the ACC in its expanded form as

$$\mathrm{ACC} = \cos\phi = \frac{\mathbf{d}_f^* \cdot \mathbf{d}_t^*}{a_t^* a_f^*} = \frac{1}{a_t^* a_f^*} \sum_{i=0}^{n} w_i \left(x_{f,i} - x_{c,i} - m_{fc}\right)\left(x_{t,i} - x_{c,i} - m_{tc}\right) \qquad \text{(A40)}$$

In the case of equal weighting of $w_i = \frac{1}{\sqrt{n}}$, these equations exactly recover the decomposition in Murphy and Epstein (1989). However, note that $(\mathrm{RMSE})^2 = (\mathrm{STDE})^2 + (\mathrm{MEANE})^2$ (as derived above, but applied to the error E) is required to put this into their exact form. The versions of STDE, SDAV, SDAF and ACC with area weighting are also clearly identical to the form being used inside Iver, as presented at the beginning of this appendix. The new forecast scores based on noise and information can similarly be derived from the new vector triangle, as for example

$$\mathrm{FI} = \mathrm{SDAF} \,/\, \mathrm{SDAV} \times \mathrm{ACC} \qquad \text{(A41)}$$

and this can be expanded to recover the summation formula given earlier in this appendix. Note that this also implies that the IE and NE, when debiased, are an orthogonal decomposition of the error standard deviation STDE. We have not explicitly distinguished the notation here in the same way as RACC vs ACC, but rather IE and NE are understood in the context of whether ACC is computed with or without subtracting the mean anomalies.

Appendix B

This appendix briefly discusses some of the edge cases that help understand the precise definitions of forecast information (FI), noise error (NE) and information error (IE), and for raising the question of these edge cases, we are indebted to the internal review by Martin Leutbecher and Thomas Haiden (both ECMWF).

A first thing to notice is that forecast information is now defined as a rescaled version of ACC and keeps the same sign as ACC, while it was strictly positive in the original definition of FTZP24. This is done for two reasons. First, ACC values can at long forecast lead times randomly attain positive or negative values, and these changes are not a-priori predictable (i.e., they cannot be calibrated out). Secondly, a definition of FI in terms of just the magnitude of the forecast anomaly projection onto the observed anomaly would lead to an inconsistent relationship with the Information Error defined in equation 12.

A further issue is that forecast information (FI) can be larger than 1, when the forecast anomaly and the anomaly correlation are large enough. In practice, this occurs in the very early forecast range, when ACC and FI are close to 1 and small chaotic variations in the forecast could project in such a way to produce FI > 1 (see Figure 3). A forecast with information greater than 1 is clearly not a better forecast than the perfect forecast (which has FI=1). And when averaging over a series of forecasts (equation A13) a set of imperfect forecasts with for example FI=0.99, 1.01, 0.98, 1.02 would provide an apparently perfect mean score of 1. We prevent this issue by working mostly with information error and noise error defined as a Pythagoras decomposition of the RMSE or STDE, and hence always positive, and with optimality defined at zero (equation 13). The mean information error in the example above would be 0.015, clearly indicating that the illustrative series of forecasts is not perfect. Further, it would be very hard in practice to generate forecasts systematically with FI > 1, somehow making the forecast activity unrealistically high while retaining forecast skill (which in practice would likely be destroyed by the high activity). Therefore, we believe there is no likelihood of forecasters "gaming" information through adjustments in activity. In any case, as illustrated in Figure 3, it is important to monitor activity as one of the key aspects of a realistic forecast. Further, this issue is irrelevant for scores presented in terms of IE and NE.

Bibliography

Bi, K., Xie, L., Zhang, H. et al. Accurate medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538 (2023). https://doi.org/10.1038/s41586-023-06185-3

Bonavita, M. (2024). On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51, e2023GL107377. https://doi.org/10.1029/2023GL107377

Candille, G. and Talagrand, O. (2005), Evaluation of probabilistic prediction systems for a scalar variable. Q.J.R. Meteorol. Soc., 131: 2131-2150. https://doi.org/10.1256/qj.04.71

Ebert, E., L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. GÅNober, S. Joslyn, B. Brown, T. Fowler, and A. Watkins. Progress and challenges in forecast verification. Meteorological Applications, 20(2):130–139, 2013. ISSN 1469-8080. doi: 10.1002/met.1392.

ECMWF, 2024. IFS Documentation CY49R1 - Part III: Dynamics and Numerical Procedures, Available from: https://www.ecmwf.int/en/elibrary/81625-ifs-documentation-cy49r1-part-iii-dynamics-and-numerical-procedures

Feng, J., Toth, Z., Zhang, J. & Peña, M. (2024) Ensemble forecasting: A foray of dynamics into the realm of statistics. Quarterly Journal of the Royal Meteorological Society, 150(762), 2537–2560. Available from: https://doi.org/10.1002/qj.4745

Geer, A.J., 2016. Significance of changes in medium-range forecast scores. Tellus A: Dynamic Meteorology and Oceanography, 68(1), p.30229, https://doi.org/10.3402/tellusa.v68.30229

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., and F. Prates (2023) Evaluation of ECMWF forecasts, including the 2023 upgrade. ECMWF Technical Memorandum n. 911. Doi:10.21957/d47ba5263c. Available at https://www.ecmwf.int/en/elibrary/81389-evaluation-ecmwf-forecasts-including-2023-upgrade

Healy, S., Bormann, N., Geer, A., Holm, E., Ingleby, N. B., Lean, K., Lonitz, K., and Lupu, C. (2024) Methods for assessing the impact of current and future components of the global observing system. ECMWF Technical Memorandum n. 916., https://doi.org/10.21957/2f240fe55f

Hoffman, Ross N. Zheng Liu, Jean-Francois Louis, and Christopher Grassoti. Distortion Representation of Forecast Errors. Monthly Weather Review, 123(9):2758–2770, September 1995. ISSN 1520-0493, 0027-0644. doi: 10.1175/1520-0493(1995)123⟨2758:DROFE⟩2.0.CO;2.

Jolliffe, I.T. and Stephenson, D.B. (2003) Forecast Verification. A Practitioner's Guide in Atmospheric Science. John Wiley & Sons Ltd., Hoboken, 240 p.

Lam, Remi and Sanchez-Gonzalez, Alvaro and Willson, et al., 2022: GraphCast: Learning skilful medium-range global weather forecasting. Arxiv preprint https://arxiv.org/abs/2212.12794

Murphy, A. H. (1973). "A new vector partition of the probability score". *Journal of Applied Meteorology.* ***12*** *(4):* 595*–600.* doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, *116*(12), 2417-2424.

Murphy, A.H. and Epstein, E.S. (1989*)*. Skill Scores and Correlation Coefficients in Model Verification. Monthly Weather Review, 117, 572-581.

Owens, R. G., Hewson, T. D. (2018). ECMWF Forecast User Guide. Reading: ECMWF. doi: 10.21957/m1cs7h

Pathak, J., S. Subramanian, P. Harrington, S. Raja, A. Chattopadhyay, M. Mardani, T. Kurth, D. Hall, Z. Li, K. Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv preprint arXiv:2202.11214, 2022

Polichtchouk, I., Stockdale, T., Bechtold, P., Diamantakis, M., Malardel, S., Sandu, I., Vana, F. and Wedi, N., 2019. *Control on stratospheric temperature in IFS: resolution and vertical advection*. European Centre for Medium-Range Weather Forecasts. doi: https://doi.org/10.21957/cz3t12t7e

Polichtchouk, I., Bechtold, P., Bonavita, M., Forbes, R., Healy, S., Hogan, R., Laloyaux, P., Rennie, M., Stockdale, T., Wedi, N. and Diamantakis, M., 2021. *Stratospheric modelling and assimilation*. European Centre for Medium-Range Weather Forecasts. doi: https://doi.org/10.21957/25hegfoq

Toth, Z., Talagrand, O. & Zhu, Y. (2005) The attributes of forecast systems: a framework for the evaluation and calibration of weather forecasts. In: Palmer, T.N. & Hagedorn, R. (Eds.) Predictability of weather and climate, Vol. 9780521848824. Cambridge University Press, pp. 584–595. Available from: https://doi.org/10.1017/CBO9780511617652.023

Wilks, D.S. (2011) Statistical Methods in the Atmospheric Sciences. 3rd Edition, Academic Press, Oxford