General description

The forecast quality assessment is described considering separately the main biases in the forecast system and a set of forecast quality metrics. The quality of the monthly averaged seasonal forecasts for the different models and variables has been assessed against ERA5 reanalysis, considering the hindcast period 1993-2016. This forecast quality assessment has been achieved through the computation of eight verification metrics: the Bias (deterministic), the Ensemble-mean Correlation (EnsCor, deterministic), the Ranked Probability Skill Score (RPSS, probabilistic), the Continuous Ranked Probability Skill Score (CRPSS, probabilistic), the Drift (deterministic), the Interquartile Range (deterministic), the Root Mean Square Skill Score (RMSSS, deterministic) and the Spread-Error Ratio (probabilistic). Last two metrics have been measured only for the teleconnection indices and replace the RPSS and the CRPSS.

Forecasts have been traditionally seen as a deterministic product, where a single response has been provided to (and sought for by) the user. In ensemble forecast systems like the ones assessed in this case the deterministic product has been formulated using the mean of the ensemble of forecasts. Its quality has been estimated using the correlation. However, in a chaotic system like climate, single deterministic forecasts can be expected to be falsified (meaning that the same exact future evolution of the climate system is not observed) by definition, no matter how good the model and initial conditions are. Indeed, when the climate system is in a particularly unstable part of the climate attractor, a forecast can be falsified dramatically by failing to predict or misplacing an extreme event completely, even when the model and initial conditions are relatively accurate. Hence, seasonal forecasts are not provided as single deterministic forecasts but the whole ensemble is used instead to formulate probability forecasts.

Probability forecasts are formulated by constructing a probability density function using a frequentist approach. These forecasts are assessed using the RPSS and CRPSS metrics. The probabilities produced by these ensembles should be reliable, which means that the events forecast occur in reality with the climatological frequency of the event. It is unrealistic to expect the raw ensemble data to be fully reliable, which penalises the metrics used and explains in part why they have lower values than those obtained with the correlation of the ensemble mean.

It is vital that sufficient hindcast data are made available to robustly estimate the quality of the systems (Hemri et al., 2020). Large hindcast samples are highly non-trivial and become computationally intensive due to the different dimensions the forecast systems include: start dates, forecast length, ensemble size and length of the hindcast period. Every time the forecast system changes, a new set of hindcasts needs to be made to provide adequate sample sizes for both the estimate of the forecast quality and provide support for the calibrations needed to take the mean biases into account and make the probability forecasts reliable.

Detailed description of the adopted metrics

Bias

Climate models exhibit systematic error (biases) due to the limited spatial resolution, simplified physics or incomplete knowledge of climate system processes. The comparison of the model's outcomes with reference datasets through the Mean Error metric helps to assess this systematic departure.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_{ref})$$

where x_i is the ensemble-mean predicted outcome in year i and x_{ref} is the observed value of the same variable in the same year i over the n years.

Ensemble-mean Correlation (EnsCor)

The Pearson correlation coefficient (Wilks, 2011) between the predicted ensemble-mean and the reference data set has been used as a measure of the linear correspondence between the retrospective predictions and the reference. This can be defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 (y_i - \bar{y})^2}}$$

where x_i and y_i are, respectively, the observed and the ensemble-mean predicted values in each season, over the $i=1,2,...,n$ years. The $\bar{x}$ and $\bar{y}$ are the average of the ensemble-mean predictions and the observations over the n years.

The EnsCor ranges between -1 and 1. If $r_{xy} = 1$ there is a perfect association between the ensemble-mean of the predictions and the observations. When $r_{xy} = 0$ indicates that there is no association between the ensemble-mean of the predictions and the reference dataset, which in turn, shows that the ensemble-mean of the predictions does not provide any added value relative to the retrospective climatology. Values of EnsCor inferior to zero ($r_{xy} < 0$) indicate that the observed climatology should be used instead of the predictions. A positive EnsCor value is the minimum requirement for seasonal predictions to have some potentially useful information because it depends not only on the potential predictability but also on the precise distribution of the data (Jolliffe and Stephenson, 2012).

Ranked Probability Skill Score (RPSS)

A comprehensive measure to evaluate the predictive skill of categorical events from probabilistic seasonal predictions is the ranked probability score (RPS; Wilks, 2011). The RPS is the sum of the squared distance between the cumulative probabilities of the *n* predictions - reference pairs (for the entire interannual series) for *k* equiprobable forecast categories (e.g. tercile):

$$RPS = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\left[\left(\sum_{j=1}^{k} y_{i,j}\right) - \left(\sum_{j=1}^{k} x_{i,j}\right)\right]^2$$

where $y_{i,j}$ and $x_{i,j}$ are, respectively, the predicted and observed probabilities assigned by the i_{th} forecast (i= 1,... n) to the k_{th} category (i= 1,... k). The $x_{i,j} = 1$ indicates that the observation is in category *k*, and $x_{i,j} = 0$ otherwise.

The RPS is often expressed as a skill score (RPSS) because it allows assessing the prediction's added value relative to the climatology. The RPSS is given by:

$$RPSS = 1 - \frac{RPS}{RPS_{clim}}$$

RPSS ranges from $-\infty$ to 1. RPSS values below 0 are defined as unskillful, those equal to 0 indicate that the forecast provides similar information than the climatological forecast, and $RPSS > 0$ shows that the predictions are better than the climatology. $RPSS = 1$ corresponds to a 'perfect' forecast.

In this assessment the RPSS has been computed for the verification of terciles (three equiprobable categories associated with the two terciles of the climatological distribution of the reference). The probabilities have been computed as the fraction of ensemble members in the corresponding category.

Continuous Ranked Probability Skill Score (CRPSS)

The continuous ranked probability skill score (CRPSS) is a commonly used probabilistic skill score that allows the predictive skill assessment of the full probability distribution (Jolliffe and Stephenson, 2012). It is based on the continuous ranked probability score (CRPS), a score that reduces to the mean absolute error if a deterministic forecast is used (Wilks, 2011). CRPS can be expressed as:

$$CRPS = \int_{-\infty}^{\infty} [F(y) - F_0(y)]^2 \, dy$$

where $F(y)$ is the cumulative density function of the predictions and $F_0(y)$ is the cumulative step function that jumps from 0 to 1 at the point where the forecast variable (y) equals to the observation (x):

$$F_0 = \begin{cases} 0, y < x \\ 1, y \geq x \end{cases}$$

The CRPS measures the difference between the predicted and observed cumulative distributions and it can be converted into a skill score (CRPSS), measuring the performance of a forecast relative to the climatology:

$$CRPSS = 1 - \frac{CRPS}{CRPS_{clim}}$$

The CRPSS ranges between −∞ to 1. CRPSS values below 0 are defined as unskillful, those equal to 0 indicate that the forecast is similar to the climatology forecast, and CRPSS > 0 shows that the predictions are better than the climatology. CRPSS = 1 indicates a 'perfect' forecast.

Root Mean Square Skill Score (RMSSS)

The RMSSS is a deterministic skill score based on the Root Mean Square Error of the ensemble-mean (RMSE). The RMSE is a deterministic measure of forecast accuracy, defined as the square root of the sum of the squared distances between the ensemble-mean predictions and observations of the reference dataset. The RMSSS is defined as 1 minus the ratio between the RMSE and the RMSE of observed climatological predictions:

$$RMSSS = 1 - \frac{RMSE}{RMSE_{clim}}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - x_i)^2}{N}}$$

Being y_i and x_i respectively the predicted and observed value for time interval *i* and *N* the total number of intervals. The RMSSS ranges from -∞ to 1. RMSSS values greater than 0 indicate that predictions are better than climatology. RPSSS = 1 corresponds to a 'perfect' forecast.

Spread-Error ratio (SoE)

The spread-over-error ratio (SoE) is a probabilistic score defined as 1 minus the ratio between the Ensemble Spread and the Root Mean Square Error of the ensemble-mean (RMSE):

$$SoE = 1 - \frac{Ens.Spread}{RMSE}$$

The Ensemble Spread is defined as the square root of the ensemble variance (the variance of the ensemble mean):

$$Ens.Spread = \sqrt{Var(y_i^m - \bar{y}_i)}$$

with *i* the time interval, *m* the ensemble member, and $\bar{y}$ the ensemble mean. SoE is a measure of reliability of the seasonal forecast systems, as predictions tend to display underdispersion, indicated by Ensemble Spread values lower than those of the RMSE (SoE lower than 1). SoE values of 1 mean that the model has the same dispersion (spread) of the RMSE.

Drift

The drift measures the steady increase or decrease of the mean spatial bias above described along the forecast times. It is computed for each start date independently by means of a simple linear regression. It is also tested for significance with a t-student test.

Operations performed to the dataset for the evaluation

This part describes the procedures and software used to process the data for the independent assessment. The EQC team has downloaded the monthly data in GRIB format with the CDS API, converted it to netcdf (e.g. with Xarray or CDO), and computed the following metrics with a specific python-based software package developed for this purpose: Bias, Ensemble-mean Correlation, Ranked Probability Skill Score (RPSS), Continuous Ranked Probability Skill Score (CRPSS), Root Mean Square Skill Score (RMSSS), Spread-Error Ratio (SoE), Interquartile Range (IQR) and Drift. The software uses the Xarray package to facilitate multidimensional array manipulation and it allows efficient parallel computing, by integrating the Dask library. Reference data is interpolated to the seasonal forecast model grid with a bilinear interpolation. For RPSS, CRPSS, RMSSS and SoE metrics, the R package SpecsVerification has been used, which is a specific library for seasonal forecast verification. This package is wrapped using the python R2Py package. The metrics were computed on ECMWF virtual machines dedicated to the evaluation and quality control. The code is available upon request at copernicus-support@ecmwf.int.

References

Hemri, S., Bhend, J., Liniger, M. A., Manzanas, R., Siegert, S., Stephenson, D. B., Gutiérrez, J. M., Brookshaw, A., & Doblas-Reyes, F. J. (2020). How to create an operational multi-model of seasonal forecasts? Climate Dynamics, 55(5), 1141–1157. https://doi.org/10.1007/s00382-020-05314-2

Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (2nd ed.). John Wiley & Sons.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences* (3rd ed., Vol. 100). Academic Press.