This paper analyses the skill over Europe of energy relevant variables from decadal prediction systems, for forecast years 1-3. It shows in which regions and seasons the variables are skilful, and compares the skill to uninitialised climate simulations to show where initialisation can enhance or reduce the skill.

Although other studies of predictive skill of energy relevant variables on decadal timescales are beginning to emerge, to my knowledge the results presented for this particular forecast timescale (<5 years) are novel, as well as those for the compound indicator. However, I have a concern that the difference in ensemble size between DCPP and HIST simulations may be contributing to the enhanced DCPP skill in regions where HIST already has skill, rather than it being solely due to initialisation. If the authors can address this, as well as a few other issues outlined below, I think the study will be worthy of publication.

General comments:

1. As mentioned above, my main concern is that the interpretation of differences in skill between DCPP and HIST simulations is due to initialisation could be hampered by the different ensemble sizes: The DCPP ensemble is more than twice as large as the HIST one, and should therefore have a better representation of the forced signal. One way to test this could be to randomly sample DCPP sub-ensembles that are the same size as the HIST one and see the effect on the results.

We thank the reviewer for this important comment. We agree that the larger ensemble size of DCPP may influence the comparison with HIST by reducing internal variability and improving the representation of the forced signal. To assess this, we will perform a sensitivity analysis based on random subsampling of the DCPP ensemble. Specifically, we will generate multiple DCPP sub-ensembles with the same number of members as HIST, recompute the ensemble mean for each subsample, and recalculate the ACC and ResCorr metrics against ERA5. This will allow us to quantify the extent to which the differences between DCPP and HIST are sensitive to ensemble size, and whether the regions of enhanced DCPP skill remain robust when both ensembles have comparable sizes. We will include the results of this analysis in the supplementary material of the revised manuscript.

2. The results section is hard to follow. The descriptions of the figures are long and detailed, and in some cases may be over-interpreting noise. For example L236: “In SON, ResCorr values are mostly positive but small, suggesting that initialization provides a limited additional contribution compared to the forced trend” – in this case the differences are statistically insignificant, so I don’t think you can infer anything here. Overall I suggest shortening the descriptions to bring out the main messages.

We thank the reviewer for this comment. We agree that parts of the Results section are overly detailed and may lead to overinterpretation, particularly in cases where signals are not statistically significant. We will revise the text to focus on the main messages, removing or simplifying sentences that describe weak or non-significant patterns, including the example highlighted. Overall, the Results section will be shortened to improve clarity and readability.

3. The reason for focussing on the 3 year timescale (as opposed to other <10 year timescales) is not explained. Is that timescale of particular interest to the energy community?

We focused on forecast years 1–3 because these lead times are particularly relevant for short- to medium-term energy planning applications, including maintenance scheduling, estimation of expected energy production, and anticipation of changes in the seasonal energy mix. In addition, this period corresponds to the lead times at which initialization is generally expected to provide the largest added value relative to non-initialized historical simulations, making it particularly suitable for assessing the contribution of initialization beyond the externally forced climate signal. The next sentence will be added to the manuscript:

- Page 2, line 52: “This 3-year forecast period is particularly relevant for short- to medium-term energy planning applications, including maintenance scheduling, estimation of expected renewable energy production, and management of seasonal energy mixes.”

Minor comments

1. L65: The paper quotes “five different decadal forecasting systems”, but it looks like only three from table S1. What’s the difference between EC-Earth3 i1, i2 and i4?

We thank the reviewer for this comment. The five forecast systems correspond to separate forecast configurations within DCPP. Three of them are based on the EC-Earth3 model, but differ in their initialization strategies (i1, i2 and i4), which are treated as separate forecast systems within DCPP.

We will revise the manuscript to clarify this point and explicitly state the ensemble composition in the main text.

- Page 3, line 66: “These forecast systems are the EC-Earth3 (with three different initialisation strategies), IPSL-CM6A-LR and MPI-ESM1.2-HR.”

2. L83: I think it would make more sense to regrid to lowest resolution, as models will not be able to predict features below this spatial scale.

In this study, the model outputs are statistically downscaled to the ERA5 grid to ensure consistency with the reference dataset used for both calibration and evaluation, and to derive impact-relevant indicators at a spatial scale relevant for applications.

We acknowledge that the native resolution of the prediction systems is coarser, and that statistical downscaling does not add intrinsic predictive skill or resolve small-scale variability. Therefore, spatially scattered or small-scale areas of significant skill should be interpreted with caution, as they may arise from the downscaling procedure or sampling variability rather than robust predictive signals. The next text will be added: “Small-scale and spatially isolated areas of significant skill are interpreted with caution, as they may arise from the downscaling procedure or sampling variability and do not necessarily indicate robust predictive signals.” (Page 6, line 172)

To minimise the risk of introducing artificial skill, we adopt a simple approach based on interpolation combined with quantile mapping. This choice is supported by previous studies (e.g., Manzanas et al., 2018; Gutiérrez et al., 2019; Moreno-Montes et al., 2026) showing that more complex downscaling techniques do not systematically provide additional predictive skill compared to simpler methods.

While regridding all datasets to a coarser resolution would provide a scale-consistent comparison, it would also smooth part of the spatial variability of the observational reference. Given the impact-oriented focus of this study, we opted for a downscaling approach while explicitly accounting for these limitations in the interpretation. We will clarify this point in the revised manuscript.

In addition, we will assess the sensitivity of the results to spatial resolution by repeating the main analysis at a coarser grid (e.g., the coarsest model resolution), in order to provide a consistent benchmark. These additional analyses are currently being performed.

- Page 3, line 83: "To provide information at a spatial scale relevant for impact-oriented applications, a statistical downscaling approach is applied to all model simulations, bringing them to the ERA5 grid (0.25°). This downscaling consists of a combination of spatial interpolation and calibration using Empirical Quantile Mapping (EQM), described below. This approach provides higher-resolution information suitable for regional decision-making processes. A simple method is adopted, as previous studies have shown that more complex downscaling techniques often provide limited added value in terms of predictive skill compared to simpler approaches (e.g., Manzanas et al., 2018; Gutiérrez et al., 2019; Moreno-Montes et al., 2026), while minimising the risk of introducing artificial skill. After interpolation and before calibration, to calculate wind energy indicators…"

3. L90: "Figure S1 shows the climatological ratio between ERA5 and GWA mean 6-hWIND over 1961–2019, which is applied as a pointwise correction to the ERA5 data". I didn't understand how you've applied the correction – do you multiply each 6h WIND at each grid point by this ratio? Does the variability look correct after this?

The correction is applied as a multiplicative factor at each grid point, computed as the ratio between the climatological mean wind speed from GWA and ERA5. Specifically, each 6-hourly ERA5 wind speed value is multiplied by this ratio, so that the mean state is adjusted while preserving the temporal variability of the original ERA5 time series. Since the same scaling factor is applied to all time steps at a given grid point, the temporal structure and variability (e.g. multiannual variability and extremes) remain unchanged apart from a proportional rescaling. This approach is used to adjust mean biases without altering the underlying variability. We will clarify this in the manuscript:

- Page 4, line 90: "Figure S1 shows the climatological ratio between ERA5 and GWA mean 6-hWIND over 1961-2019, which is applied as a multiplicative pointwise correction factor (GWA/ERA5) to adjust the mean wind speed while preserving the temporal variability of the ERA5 data."

4. L94-102: It is unclear to me how the EQM is performed. For the historical simulations, if you take per calendar year per member, do you take, for example, simulated year 1990 from member 1 and map the quantiles to 1990 in the reference dataset (corrected ERA5)? If so, as 1990 was a very windy year in some regions, will that inflate the skill? You mention cross-validation, but if you remove the year in question, which years do you use as the reference? Also, is EQM performed at each grid point?

EQM is applied independently at each grid point using a leave-one-year-out cross-validation approach. For a given target year (e.g. 1990), the empirical distributions used to estimate the quantiles are constructed excluding that year, using all remaining years in the period within a moving temporal window around each target day of 31 days. The modelled values for the excluded year are then bias-corrected using the quantile mapping derived from this reference distribution, ensuring that no information from the target year is used in the calibration. Therefore, anomalous conditions in the target year do not influence the estimation of the quantile mapping transfer function, avoiding artificial skill inflation associated with in-sample correction.

This procedure is applied independently at each grid point and per member (and per forecast year in the case of decadal predictions to account for the model drift). We will clarify this in the manuscript.

- Page 4, line 99: “A leave-one-year-out cross-validation is used to prevent overfitting and artificial skill inflation (Elsner and Schmertmann, 1994). For each target year, the empirical distributions used for the quantile mapping are constructed excluding that year, using all remaining years in the period. The calibration is applied independently at each grid point.”

5. L206-215: comparison of ResCorr to HIST for PVpot (and where it appears in other sections) – it’s hard to compare across figures in the main paper and supplementary material. I suggest either including HIST ACC plots in the main paper as additional panels, or you could show contours where HIST has significant skill (e.g. on Fig 1f-j).

We agree that the comparison between DCPP and HIST skill is not straightforward when the information is split between the main figures and the supplementary material. To improve clarity, we will revise the figures to facilitate a more direct comparison, for example by including additional visual information on the HIST skill in the main panels. This will allow a clearer assessment of the differences between DCPP and HIST throughout the manuscript.

6. Fig 2 – It would help to have titles on each panel. This also applies to Fig 4 and 6 where the panel titles aren’t given in the caption.

The titles of each panel will be included in Figures 2, 4, and 6.

Technical/typos:

1. There seems to be an issue with some of the referencing. For example L21 “Agency, 2022”, L30 “Commission, 2020”; L34 “Association et al. 2023”.

We agree that the referencing of institutional sources was unclear. We will revise these citations to explicitly include the full name of the institutions (e.g. European Environment Agency, European Commission) to ensure proper attribution and consistency.

2. L168 “the ACC between the DCPP of the indicator and ERA5” - do you mean “the ACC of the indicator between DCPP and ERA5”? Similar phrases appear elsewhere, e.g. L174.

We agree that the phrasing was unclear. We will revise the text to improve clarity and ensure that it consistently refers to the ACC of the indicator computed between DCPP and ERA5.

- Page 6, line 168: “In each section, the ACC of the indicator between DCPP and ERA5 and the ResCorr between DCPP and ERA5 relative to HIST are shown for the annual mean and individual seasons.”

- Page 6, line 174: “The ACC between DCPP and ERA5 for the variables used to calculate the indicators (RSDS, TAS and SFCWIND) is shown in Figure S5.”

- Page 7, caption Figure 1: “ACC between DCPP and ERA5 for solar PVpot for the annual mean…”

- Page 14, line 360: “Figure 5a–e shows the ACC between DCPP and ERA5 for NED”

3. In the interannual variability plots (Figs S6, S9, S11) – is it interannual variability, or variability of 3-year means?

The variability shown in Figures S6, S9 and S11 is computed from time series smoothed using a 3-year rolling mean, consistent with the forecast period (years 1–3). The standard deviation is then calculated from these smoothed series. Therefore, the metric reflects multiannual variability at the 3-year timescale rather than strictly interannual variability. We will revise the manuscript accordingly to avoid confusion.

Therefore, the metric represents the variability of 3-year mean values rather than strictly interannual variability. We will revise the manuscript to clarify this point and avoid potential confusion.

- “The associated multiannual standard deviation is low across the domain (S6f–j), indicating that PVpot variability is small and largely controlled by mean climatological conditions.”

- Page 10, line 256: “Figure S9 shows the climatology (Figure S9a–e) and the multiannual standard deviation (Figure S9f–j) of WCF for the annual and seasonal means.”

- Page 10, line 262: “Unlike PVpot, WCF exhibits a non-negligible multiannual standard deviation, with the largest values (up to 5–6 percentage points) occurring mainly in DJF and over northern Europe, reflecting stronger wind variability.”

- Page 14, line 341: “Figure S11 shows the climatology (Figure S11a–e) and the multiannual standard deviation (Figure S11f–j) of NED for the annual mean and each season.”

4. Fig S5: labelling – the top row is labelled f-j, the middle row is a-e – unclear which is TAS and RSDS.

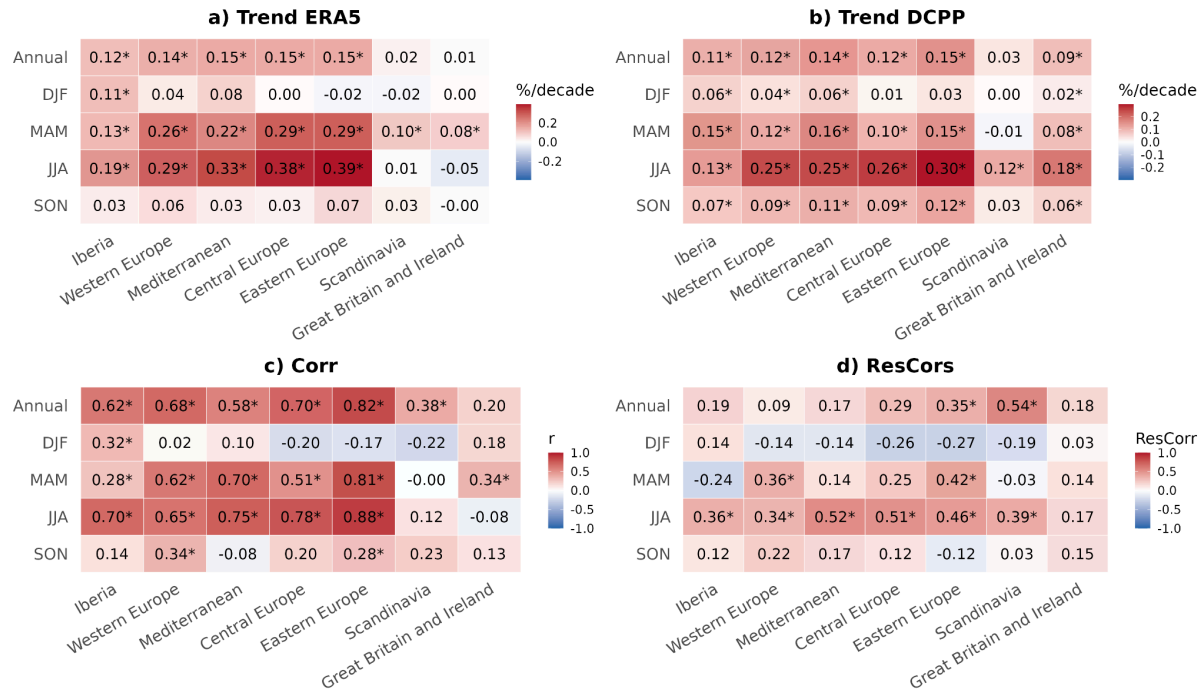
We agree that the panel labelling in Figure S5 is unclear. We will revise the figure to ensure that the ordering and labelling of the panels are consistent and that the variables (TAS and RSDS) are clearly identified.

5. Fig 2b – Scandinavia DJF has correlation of 0.00 but it is labelled as statistically significant.

In the previous version, trend significance was assessed using the standard linear regression p-value, which did not account for temporal autocorrelation. We have now revised the trend significance test and applied a modified Mann–Kendall test following Hamed and Rao (1998), which corrects the variance of the test in the presence of serial dependence.
With this revised approach, the DJF trend for Scandinavia in Figure 2b is no longer statistically significant. The significance of several other weak trends has also been reduced. We will update Figure 2, Figure 4, Figure 6 and Figure S8 accordingly and now provide the revised heatmaps, in which the trend values and significance markers reflect the autocorrelation-corrected test.

- Page 6, line 160: “Trends are estimated from the annual time series using a linear fit, while statistical significance is assessed independently using a modified Mann–Kendall test that accounts for temporal autocorrelation following Hamed and Rao (1998). This correction reduces the effective significance of weak but serially dependent trends.”

Updated Figure 2:



6. L225 “For Scandinavia and Great Britain and Ireland, trends generally disagree and correlations are low, with a few exceptions (Scandinavia in the annual mean and SON and Great Britain and Ireland in MAM), although significant correlations are limited to the the annual mean in Scandinavia and MAM in Great Britain and Ireland.” I found this confusing – isn’t the second part repeating what’s in the brackets?

We thank the reviewer for this comment. We agree that the sentence is confusing, as it combines different pieces of information. The intention was to distinguish between cases where trends have the same sign and those where this agreement results in significant correlations. We will revise the text to clarify this distinction.

- Page 8, line 221: “For Scandinavia and Great Britain and Ireland, correlations are generally low, with significant values only found for the annual mean in Scandinavia and MAM in Great Britain and Ireland, where ERA5 and DCPP exhibit consistent trend signs.”

7. Label for Fig S14 – says same as Fig S5 – do you mean S7?

We thank the reviewer for pointing this out. This is a mistake in the figure caption. It should refer to Figure S7, and we will correct it in the revised manuscript.

8. L438 “multi-model ensemble”

This is a typo, and “multi-model” should read “multi-model ensemble”. We will correct this throughout the manuscript to ensure consistent terminology.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., ... & Pagé, C. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross‐validation experiment. International journal of climatology, 39(9), 3750-3785.

Hamed, K. H., & Rao, A. R. (1998). A modified Mann-Kendall trend test for autocorrelated data. Journal of hydrology, 204(1-4), 182-196.

Manzanas, R., Gutiérrez, J. M., Fernández, J., Van Meijgaard, E., Calmanti, S., Magariño, M. E., ... & Herrera, S. (2018). Dynamical and statistical downscaling of seasonal temperature forecasts in Europe: Added value for user applications. Climate Services, 9, 44-56.

Moreno-Montes, S., Delgado-Torres, C., Duzenli, E., Pérez-Zanón, N., Marcos-Matamoros, R., & Soret, A. (2026). Comparative analysis of statistical downscaling methods for multi-model decadal climate predictions over Western Europe. Climate Services, 42, 100639.