

Moreno-Montes et al 2026 Reviewer #1 Comments

Title: Decadal predictions of wind, solar and compound power indicators to support the European renewable energy sector

Author(s): Sara Moreno-Montes et al.

MS No.: egusphere-2026-1205

MS type: Research article

General Comments

This manuscript evaluates the ability of a multi-model hindcast of decadal predictions to forecast energy indices across Europe for forecast years 1-3. The authors report some skill for forecasts of solar PV potential, while more limited skill is present for wind capacity factors. They also construct a compound energy drought indicator and identify seasonally dependent prediction skill. While the approach presented here is novel, there are several issues with the methodological approach and interpretation of the results which may delay publication until further evaluation has been conducted.

We sincerely thank the reviewer for the thorough and constructive evaluation of our study, which will certainly improve the quality and readability of the manuscript. We appreciate the detailed comments provided, which have helped us identify several important methodological and interpretative aspects that required clarification or further analysis. Point-to-point responses to each comment (in blue) are presented below.

The decision to regrid the decadal predictions (~1 degree resolution) to the higher resolution of ERA5 (0.25 degree) is questionable here, as this gives the impression that decadal predictions have the ability to capture small scale variability in specific regions. This makes it more difficult to interpret the skill maps presented, as much of the significant positive skill is demonstrated in scattered, small scale areas. It would be more appropriate to regrid ERA5 and the decadal prediction systems to the resolution of the coarsest model used (e.g., 1.25 degrees) to provide a fair comparison of skill on the appropriate spatial scales of the decadal predictions. Especially as much of the discussion is focused on slight changes in skill over small regions, a fairer comparison, like the coarse regridding suggested, would help to clarify signal from noise. This would help the interpretation of the findings, as the areas which do show significant skill should be visible as larger regions. This would further demonstrate the robustness of the findings presented here.

We fully acknowledge that statistical downscaling has well-known limitations, including its reliance on empirical relationships, assumptions of stationarity, and the fact that it does not increase the intrinsic predictability of the system (Maraun et al., 2010; Maraun & Widmann, 2018; Hewitson et al., 2014). Despite these limitations, statistical downscaling techniques are widely used in the context of climate services to provide climate information at spatial scales relevant for end users and sectoral applications (e.g., energy, water management),

rather than to enhance dynamical skill (Maraun et al., 2010; WMO, 2011; Soares et al., 2017). In this framework, high-resolution information is often required to derive impact-relevant indicators.

To increase the spatial resolution of the predictions while preserving the variability of the observation-based reference dataset, we adopted a simple approach based on interpolation combined with quantile mapping. Quantile mapping corrects systematic biases while preserving the non-Gaussian characteristics of wind distributions, which are essential for the indicators analysed here. This choice is supported by studies showing that more complex downscaling methods do not systematically outperform simpler approaches in terms of skill (e.g., Manzananas et al., 2018; Gutiérrez et al., 2019; Moreno-Montes et al., 2026).

However, we agree that spatial features at scales smaller than the native resolution of the prediction systems should be interpreted with caution, as they may arise from the downscaling procedure or sampling variability rather than robust predictive skill. In the revised manuscript, we assess the sensitivity of the results at a coarser grid (e.g., the coarsest model resolution) to evaluate the robustness of the findings. These additional analyses are added to the Supplementary Material and all the results obtained are compared with them in the main text.

The next sentences have been added:

Page 4, lines 93-95: “To assess the sensitivity of the results to the spatial resolution used for the analysis, all indicators and skill metrics are additionally computed on the native spatial resolution of the coarsest prediction system considered (IPSL-CM6A-LR: $2.5^\circ \times 1.25^\circ$).”

Page 7, lines 218-220: “Similar skill patterns are obtained at the coarser resolution (Figure S6a-e), indicating that the main results are robust to spatial resolution. “

Page 9, lines 246-247: “Similar conclusions are obtained at the native resolution of the coarsest prediction system (Figure S6f-o), where both the ACC of HIST and the ResCorr show broadly consistent patterns.”

Page 11, lines 295-296: “These regions of significant skill are also present in the coarsest-resolution results (Figure S11a-e).”

Page 13, lines 321-322: “Similar conclusions are obtained at the native resolution of the coarsest prediction system (Figure S11), which shows the same large-scale patterns, supporting the robustness of these results.”

Page 15, lines 373-374: “Overall, the skill patterns are broadly consistent with those obtained at the coarsest resolution (Figure S15a-e). “

Page 15, lines 392-396: “Similar conclusions are obtained at the native resolution of the coarsest prediction system (Figure S16f-o), with broadly consistent ACC and ResCorr patterns. The main exception is DJF (Figure S16g), where HIST exhibits substantially fewer significant ACC values at the coarser resolution. This suggests that part of the additional skill identified in the high-resolution results during this season may reflect spatial noise rather than robust large-scale predictability.”

The density of figures and subplots described in the results section makes the findings of the paper hard to follow. For example, on page 7 the results of Figure 1 (10 panels) are described with reference to Figure S5 (15 panels) in the supplementary information. In order to make the findings of this paper more impactful, it might be appropriate to condense the findings described in the results section and reduce references to the supplementary material to only where necessary. This would help the reader to better understand the main findings of the paper and improve the clarity of the results.

We thank the reviewer for this helpful comment. We agree that the current density of figures and cross-references to the supplementary material may difficult the readability of the results section. In the revised manuscript, we have streamlined the presentation by condensing the main findings, reducing the number of panels where possible, and limiting references to supplementary figures to cases where they are strictly necessary. In particular, we have included the HIST ACC results in Figures 1, 3 and 5.

The authors report that the main contributing factor to the number of energy drought days (NED) during the winter (DJF) is the number of effective wind generation days (wind-Neff). However, in the supplementary material (Figure S13b, i) they demonstrate that solar-Neff values are missing across most of Europe (aside from southern Iberia). This means that the number of ineffective solar days component of the NED index is effectively constant. The authors repeatedly claim that the areas of NED skill during DJF primarily align with wind-Neff. While this is correct, this is also misleading, as for most of the area shown in Figure 5b and 5g, only the number of ineffective wind days contributes, so the NED shown is no longer a compound index. For clarity, the authors should consider presenting the results differently for DJF in Figure 5 or rephrasing the findings to be clear about what is shown.

We agree that during DJF the solar component of the NED index shows very limited variability over most of Europe, as solar-Neff is effectively zero in large regions. As a consequence, the variability of NED in this season is primarily controlled by the wind component. However, NED remains a compound indicator by definition, as it is based on the joint occurrence of ineffective wind and solar conditions. In winter, the lack of effective solar production means that this condition is always fulfilled, so NED effectively reflects the occurrence of ineffective wind days. This does not imply a change in the definition of the index, but rather a seasonal regime in which one component does not contribute to variability. We have revised the manuscript accordingly.

- Page 15, line 383-386: “Over most of Europe in DJF and in northern regions in SON, as solar-Neff remains close to zero (Figure S14b,e), ACC can not be computed (Figure S16b,e) and NED ACC is directly the wind-Neff one (Figure S16g,j). In SON, outside these regions, the weak NED skill is more consistent with wind-Neff (Figure S16j), although some contribution from solar-Neff (Figure S16e) is also apparent.”

Overall, this paper applies a novel methodology to demonstrate some notable skill for energy-sector metrics over Europe. However, methodological choices currently limit the robustness of the results presented. Until these issues are addressed, principally that of fair model comparison (coarse resolution), publication should be delayed. To allow time for these points to be addressed, I recommend that publication to ESD is declined until these issues can be addressed.

We thank the reviewer for the positive assessment of the novelty and relevance of our work, as well as for the constructive comments provided. We have carefully considered all the concerns raised, particularly those regarding the comparison at the coarse model resolution. In response, we have performed the additional coarse-resolution analysis suggested by the reviewer and incorporated it into the revised manuscript and Supplementary Material. We believe that these additions, together with the methodological clarifications and the more cautious interpretation of the results, have substantially strengthened the manuscript and addressed the reviewer's concerns.

Specific Comments

- Pg3, lines 65-66: "In this study, five different decadal forecast systems contributing to the Decadal Climate Prediction Project (DCPP; Boer et al., 2016) of the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016), are used."
 - Technically, only three different decadal forecast systems are used, as 3 of the 5 forecast system are different initialisations of EC-Earth3. Although the ensemble size is detailed in the SM, it would be useful to mention this here. Has the drift of each of the models been corrected here, as in Appendix E of Boer et al., 2016 (10.5194/gmd-9-3751-2016)?
- We agree that the description of the forecast systems may be ambiguous and that the ensemble composition should be more clearly stated in the main text. While the five forecast systems correspond to distinct DCPP contributions, three of them are based on the EC-Earth3 model but differ in their initialization strategies and experimental configurations. We have revised the manuscript to clarify this point and explicitly state the ensemble composition.
 - Page 3, lines 69-70: "These forecast systems are the EC-Earth3 (with three different initialisation strategies), IPSL-CM6A-LR and MPI-ESM1.2-HR."
- We thank the reviewer for raising this point. In the revised manuscript, model drift has been addressed by applying drift correction prior to calibration. Specifically, mean model outputs have been adjusted to the observed climatology by removing the model mean and replacing it with the corresponding observed mean, both computed over a common reference period defined by the forecast configuration. Given that the analysis uses start dates from 1960 to 2016 and forecast years 1–3, the common period across all lead times spans from 1963 to 2017. In addition, to ensure consistency with the cross-validation framework used in the calibration, the climatological means have been computed excluding the target year being calibrated. That is, for a given year, the reference climatology is estimated using all other years within the 1963–2017 period, except the year under consideration. This approach will reduce the impact of model drift while preserving the temporal variability of the simulations and avoids introducing artificial skill. The Empirical Quantile Mapping (EQM) calibration has then been applied to these drift-adjusted time series.
 - Page 4, lines 104-109: "Once wind-specific corrections have been applied, model outputs are adjusted to the observed climatology by removing the model mean and replacing it with the corresponding observed mean. These climatologies are computed over a common reference period (1963–2017),

defined by the availability of all forecast years included in this study (i.e. 1–3) across the selected start dates (1960–2016). To ensure consistency with the cross-validation framework, the climatological means are estimated excluding the target year being calibrated. This adjustment reduces the impact of model drift while preserving the temporal variability of the simulations (Boer et al., 2016).”

- Pg 3, lines 67-68: “the indicators are computed for the three corresponding CMIP6 non-initialized historical forcing simulation models.”
 - It would also be useful to include the number of members here.
- The number of members have been included in the main text:
 - Page 3, lines 71-72: “the indicators are computed for the three corresponding CMIP6 non-initialized historical forcing simulation models (7 EC-Earth3 members, 11 IPSL-CM6A-LR members, and 2 MPI-ESM1.2-HR members)”
- Pg 3, line 83: “To ensure consistency across datasets, variables from all model simulations are first interpolated to the ERA5 grid (0.25°).”
 - Can you explain why you chose to regrid the model simulations to a higher resolution for comparison, rather than regridding the ERA5 data to a constant model resolution (e.g., all models and ERA5 data compared at 2.5 x 2.5 degrees)? No new information is gained from the models from regridding to the higher resolution of ERA5. It would be useful to know how much the results of this study would change, if all analysis was done with models and ERA5 at a 1.25 degree (or 2.5 degree) resolution, to ensure a fair comparison.
- Renewable energy users currently rely on reanalysis products for their applications (Staffel et al., 2016; Dubus et al., 2023; Chen et al., 2025); therefore, it is essential that decadal predictions are provided at the same spatial resolution. This represents a fundamental requirement for the provision of climate services, as these users often lack the computational capacity and technical resources needed to further downscale or increase the spatial resolution of decadal prediction products themselves. We acknowledge that this downscaling procedure does not add new information from the models and may give the impression of resolving finer-scale variability than is intrinsically represented. As discussed above, we have clarified this point in the manuscript.
 - Page 3, lines 88-93: “To provide information at a spatial scale relevant for regional applications, a statistical downscaling approach is applied to model simulations to increase the horizontal resolution, bringing them to the ERA5 grid (0.25°). This downscaling procedure consists of a combination of spatial interpolation and calibration using Empirical Quantile Mapping (EQM), described below. This 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.”

In addition, we have assessed 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.

- Pg 4, lines 90-91: “In addition, ERA5 wind speeds are adjusted using the Global Wind Atlas (GWA) to better represent long-term mean conditions at hub height.”
 - I am not clear here as to whether the 100m wind speeds are used from ERA5 or the 10m wind speeds passed through the power law transform. It would be useful to be clearer with this. In the SM, it appears that 10m wind speeds were used from ERA5 and then extrapolated using the power law up to 100m. How would the results change if 100m wind speeds from ERA5 were used instead?
- We confirm that 10 m wind speeds from ERA5 are used and extrapolated to hub height using a power law formulation, ensuring methodological consistency with the model outputs, which do not provide native wind speeds at 100 m. Applying the same vertical extrapolation framework to both observations and model simulations avoids introducing inconsistencies associated with comparing differently derived wind products.

Using native ERA5 100 m wind speeds would nevertheless represent an alternative observational reference. Native ERA5 100 m wind speeds are diagnostically derived within the IFS boundary-layer scheme using stability- and roughness-dependent formulations, rather than being directly simulated variables (Peña and Gryning, 2008). Consequently, ERA5 native 100 m winds and the power-law-derived winds represent different hub-height formulations. Following the reviewer’s suggestion, we compared the 1961–2019 climatologies obtained from ERA5 10 m winds extrapolated to 100 m against native ERA5 100 m winds (Figure A1). The figure shows the difference between the power-law-derived and native ERA5 100 m climatologies. Results reveal spatially coherent differences, with predominantly negative values over land and positive values over ocean regions, indicating that the power-law approach tends to underestimate mean hub-height winds over land and overestimate them over the ocean relative to ERA5 100 m winds. Consequently, using native ERA5 100 m winds would modify the mean state of the wind-based indicators and associated capacity factors. Although differences in the mean state are evident, the impact on predictive skill is expected to be more limited, since the evaluation primarily focuses on temporal variability and relative skill within a consistent ERA5–DCPP framework.

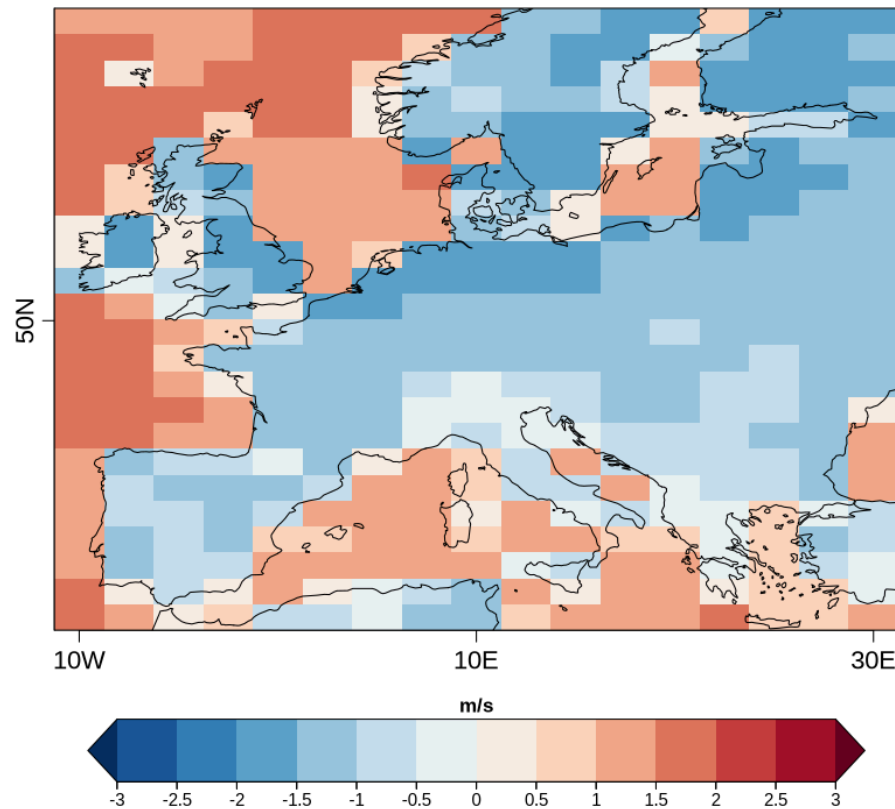


Figure A1. Difference between the 1961–2019 climatological mean 100 m wind speeds derived from ERA5 10 m winds using the power-law extrapolation and native ERA5 100 m wind speeds.

- Pg 4, lines 91-92: “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.”
 - With this pointwise bias correction of ERA5 to the global wind atlas, is this an additive (correcting the mean) or a multiplicative (correcting the spread) bias correction? It would be useful to have greater detail on how exactly this works.
- The adjustment of ERA5 wind speeds using the GWA is applied as a multiplicative correction factor, computed as the ratio between GWA and ERA5 mean wind speeds. This approach adjusts the mean state while preserving temporal variability, and we have clarified this in the manuscript:
 - Page 4, line 101-103: “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.”
- Pg 4, lines 95-96: “EQM maps the quantiles (percentiles) of the modelled data to the observed ones without assuming any underlying distribution and then it corrects the model biases.”
 - With the Empirical Quantile Mapping, it would be useful to clarify the moments of the distribution that this method acts to correct. This is currently performed for each forecast year and member, but how would it change the

results if the EQM was performed across all ensemble members for each of the models in turn? With the signal to noise issue (Scaife and Smith, 2018, 10.1038/s41612-018-0038-4), we cannot assume that the individual members can be regarded as equivalent realisations of climate variability as in the observations. Will the EQM method applied to each member in turn artificially inflate the variance? How is the signal to noise issue accounted for in this study?

- Empirical Quantile Mapping (EQM) corrects the distribution of the variable by adjusting empirical quantiles at regular intervals (here, percentiles from the 1st to the 99th, one by one). This approach allows the method to correct not only the mean and variance, but also higher-order moments such as skewness. We have clarified this in the manuscript:
 - Page 4, lines 111-112: “By adjusting empirical quantiles at 1% intervals (from the 1st to the 99th percentile), EQM corrects the distribution of the variable, including the mean, variance, and higher-order moments.”

Regarding the calibration strategy, we acknowledge that an alternative approach would be to estimate the EQM transfer function from the pooled distribution of all ensemble members for each model. In this study, the member-by-member strategy has been used because it preserves a more comparable sample size between observations and model data during calibration. To increase the robustness of the estimated empirical distributions while maintaining this consistency, the calibration was performed using a moving temporal window around each target day.

This approach is also consistent with the findings of Roberts and Leutbecher (2025), who first analysed the impact of pooled-member and member-by-member approaches on anomaly estimation, and subsequently evaluated their effect on ensemble calibration and reliability diagnostics. They showed that member-by-member anomaly estimation preserves statistically consistent spread-error relationships, whereas pooled-member approaches can introduce biases in ensemble reliability diagnostics. Although their framework differs from the present study, as it focuses on anomaly estimation and simpler calibration approaches rather than EQM-based full-distribution calibration, they similarly found that differences between pooled-member and member-by-member approaches become small when sufficient calibration data are available.

However, we acknowledge that ensemble members cannot be considered independent realisations of the observed climate and that signal-to-noise limitations are inherent to decadal prediction systems (Scaife and Smith, 2018). To assess the potential relevance of this issue in the calibrated system, we additionally evaluated the Ratio of Predictable Components (RPC; Figure A2), computed as the ratio between the predictable signal estimated from the ensemble mean and that represented by the ensemble members. In this framework, RPC values significantly larger than 1 indicate that the ensemble mean is more strongly correlated with observations than expected from the ensemble-member signal, a behaviour commonly associated with signal-to-noise issues in decadal prediction systems. Statistical significance was assessed using a 5-year block-bootstrap procedure in which start dates were resampled with replacement and the RPC was recalculated

for each resampled dataset to estimate confidence intervals. Confidence intervals were estimated from 1000 bootstrap resamples.

RPC was analysed only in regions with $ACC > 0$, since negative ACC values do not allow a meaningful interpretation of RPC. The results show that there are not generally significant RPC greater than 1 across all indicators and seasons, although some localized and statistically significant regions remain, particularly for the NED indicator during JJA. Overall, the results suggest that signal-to-noise inconsistencies may affect specific regions and seasons, but do not dominate the behaviour of the indicators over the analysed domain.

We have clarified in the revised manuscript that future studies could further investigate the signal-to-noise characteristics of the calibrated system.

Page 20, lines 491-494: "Furthermore, as decadal prediction systems are known to be affected by signal-to-noise problems (Scaife and Smith, 2018), future studies could further investigate the impact of signal-to-noise characteristics of the calibrated ensemble or apply correction methods such as the NAO-matching proposed by Smith et al. (2020)."

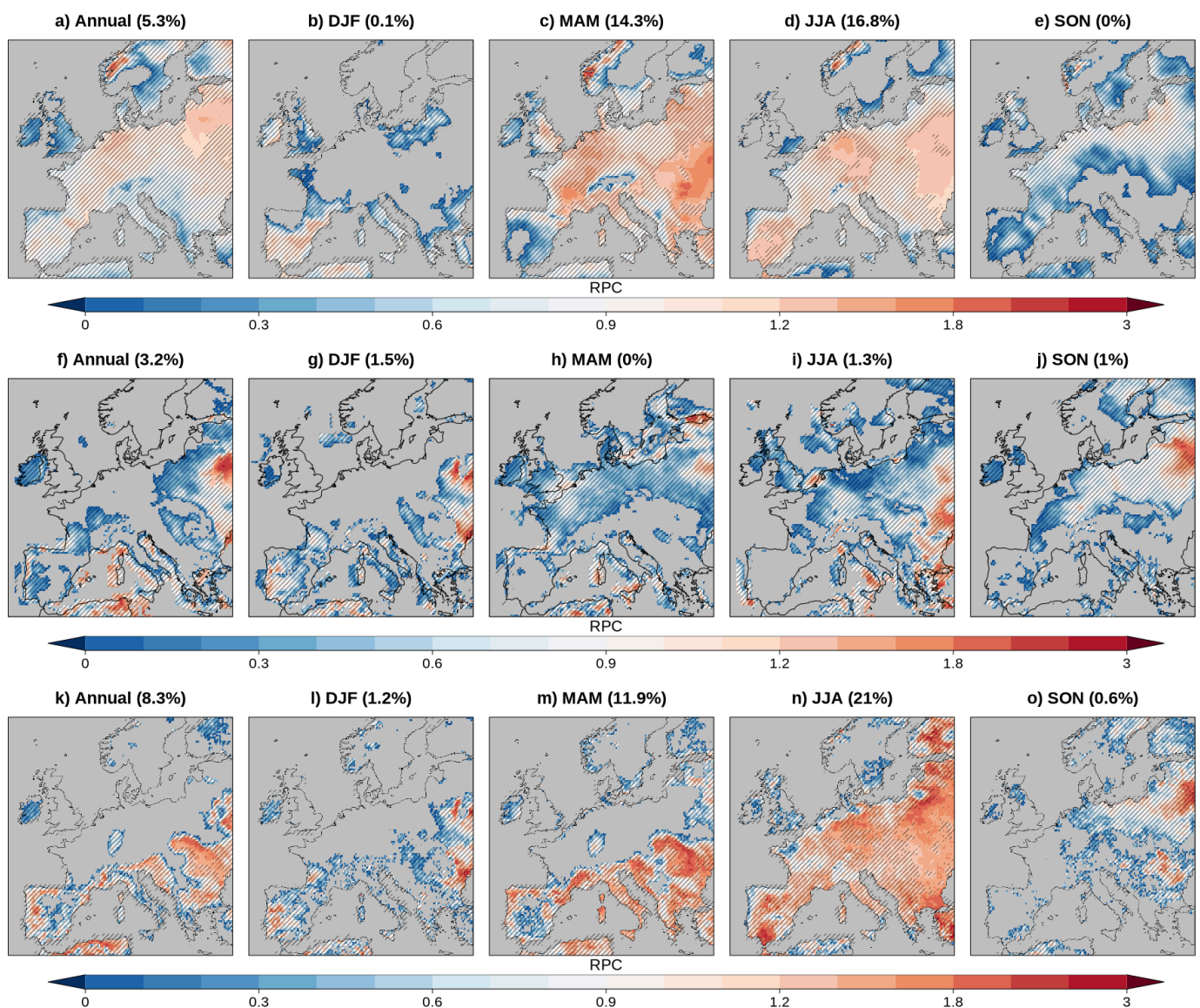


Figure A2. Ratio of Predictable Components (RPC) for PVpot (top row), WCF (middle row), and NED (bottom row) for Annual, DJF, MAM, JJA, and SON averages over forecast years 1–3. RPC was computed only in regions with positive ACC. Hatched areas indicate regions where RPC is not significantly different from 1 based on a bootstrap test using resampled start dates. Percentages in panel titles indicate the fraction of the analysed area with statistically significant RPC values.

- Pg 4 lines 111-114: “PVpot is computed as a dimensionless metric that combines RSDS, TAS and SFCWIND to represent the performance of photovoltaic cells under ambient environment. PV power generation at a specific location is determined by multiplying PVpot with the nominal installed capacity of that location.”
 - How do you determine the nominal installed capacities at each location here?
- Thank you for the comment. PVpot is a dimensionless indicator representing relative photovoltaic performance under given meteorological conditions and does not depend on prescribed installed capacities. Installed capacity would only be required to translate this indicator into absolute power generation for specific applications. Several datasets provide information on the nominal installed capacity of photovoltaic

systems, either at individual solar park locations or aggregated across administrative regions.

Global Solar Power Tracker : [Link](#)

PECD: [Link](#)

- Pg 7, Figure 1:
 - It would help the reader to have a title for the colorbar here (e.g., ACC). Also, it may help to have ylabels on the far left plots in each row, identifying “Full field” and “Residual” correlation skill. This goes for Figures 3, 5, and similar plots in the supplementary material as well (e.g., S5 onwards).
- We agree that the interpretation of the figures can be improved. Since the three rows represent different skill metrics (ACC for DCP, ACC for HIST, and ResCorr), a single colorbar title would not be fully representative of all panels. Instead, we have added row labels on the left-hand side of the figures identifying the metric shown in each row (“DCP ACC”, “HIST ACC”, and “ResCorr”). This modification has been applied consistently to Figures 1, 3, 5 and the corresponding figures in the Supplementary Material.
- Pg 7 lines 182-183: “This reduction is mainly due to northern Europe and parts of Iberia, where MAM skill is generally non-significant.”
 - Presumably due to the lack of skill in northern Europe here?
- The sentence has been changed to:
 - Page 7, lines 215-217: “In MAM (Figure 1c), the ACC is not significantly positive in northern Europe and Iberia and the percentage of significant area is lower than in the annual mean and JJA, although the skill patterns are similar.”
- Pg 8 lines 210-211: “On the other hand, DCP skill (Figure 1b) is lower than HIST skill (Figure S7b) in areas such as southern Iberia or the Alps”
 - To my eye, the DCP (Figure 1b) shows no skill over the Alps and marginal skill over eastern Spain. This marginal skill over Spain increases slightly in HIST, but the skill over the alps in Figure S7b appears too small to be significant. It might be helpful to remove the statement about the Alps here for clarity, therefore.
- The statement about the Alps has been removed and the sentence simplified to:
 - Page 9, lines 243-244: “In DJF, DCP skill (Figure 1b) is lower than HIST skill (Figure 1g) in areas such as southern Iberia.”
- Pg 9, Figure 2:
 - Can you explain how the trend of 0.00 in Figure 2b DJF Scandinavia can be statistically significant?
- Thank you for pointing this out. 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 significance of several weak trends has been reduced. We have updated 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 193-195: “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).”

As an example, the updated Figure 2:

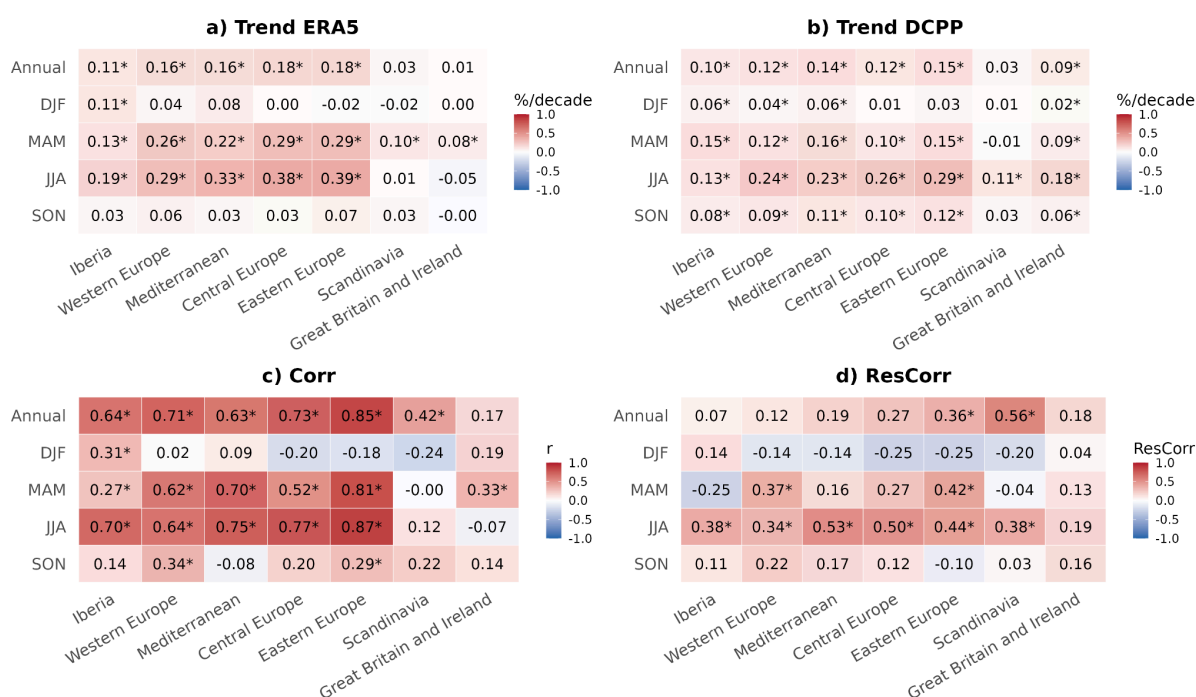


Figure 2. PVpot trends derived from ERA5 for years 1961-2019 with a three years rolling mean (a) and from the DCP for the start dates 1960-2016 for over forecast years 1–3 (b), the correlation between ERA5 and DCP (c), and the ResCorr between DCP and ERA5 respect HIST (d) shown for the annual mean and for each season across the seven european sub-regions (Figure S4). Statistically significant trends and correlations at the 95% confidence level are marked with an asterisk. Trends significance are assessed using a modified Mann–Kendall test accounting for temporal autocorrelation.

- Figures 2, 4, 6:
 - In all of these figures, it would be helpful to make sure the colorbars for trends (a and b) and correlations (c and d) are the same in each figure. It makes it more difficult to compare the trends particularly when the colorbar scales are slightly different.
- Thank you for the suggestion. The colorbars have been corrected in the revised version ranging all of them from -1 to 1.

- Pg 10, lines 248-249: “The ResCorr isolates the added value of initialization beyond the historical simulations, providing insight into the role of internally generated variability.”
 - The residual correlations show the added value of initialisation, but do not necessarily isolate the internal variability. Initialisation can improve the response to external forcing when initialisation corrects the model response to external factors (Smith et al., 2019, <https://doi.org/10.1038/s41612-019-0071-y>). Consider rephrasing this sentence accordingly.
- We agree that residual correlation does not strictly isolate internally generated variability, as initialization may also improve the response to external forcing. We have rephrased this point to provide a more accurate interpretation.
 - Page 10, lines 273-275: “The ResCorr isolates the added value of initialization beyond the historical simulations, providing insight into climate variations not captured by the historical simulations.”
- Figures 3-6:
 - It would be useful to include the captions here, rather than refer back to the previous Figure through “...Same as Figure X”. This would make it easier for the reader to interpret the findings. As mentioned above for previous figure, it would be very useful to include captions for the colorbar (e.g., ACC) and potentially ylabels identifying what is shown in each row.
- The complete captions and labels have been included in the revised manuscript.
- Page 12, lines 293-294: “For the annual mean and DJF (Figure 3f–g), initialization reduces WCF skill over Scandinavia and northern Europe, although HIST skill in these regions is already non-significant”
 - To my eye, there is no significant positive skill over Scandinavia and Northern Europe in either the DCPD or HIST results for WCF in Figure 3 or Figure S10. Suggesting that “initialisation reduced WCF skill over...” could be misleading here, as, although this statement is correct looking at Figure 3f-3g, it suggests the existence of predictable skill for WCF over Scandinavia and Northern Europe in the first place. I find it difficult to interpret the significance of initialisation reducing skill where there is none in the first place. Consider rephrasing or removing.
- We agree that this statement may be misleading in regions where skill is not statistically significant. We have therefore removed this sentence to improve clarity.
- Pg 12, line 295: “The skill reduction may be related to the high interannual variability of WCF in these regions (Figure S9f–g),”
 - In Figure S9, I find it very difficult to interpret whether the interannual standard deviation over Scandinavia and Northern Europe is higher than in other regions. To my eye, the magnitudes of interannual variability are similar in Scandinavia and Northern Europe to the rest of Europe, with the one exception to this being the Baltic region. I am not clear how the skill reduction here (i.e., less negative skill in HIST to more negative skill in DCPD) relates to interannual variability. Consider rephrasing or drop.
- We agree that this interpretation is not sufficiently supported by the figure and may be unclear. We have therefore removed this statement to improve clarity.

- Page 12, lines 302-304: “During JJA (Figure 3i), initialization increases skill over parts of eastern Europe and central Germany, while reducing it over Iberia, where HIST skill is already low (Figure S10d), resulting in a higher fraction of significant area for the DCPD than for the HIST”
 - In this case, the areas of significance over central Germany are very small so could be a result of noise. Additionally, as mentioned in the point above, I am not clear on the significance of negative skill (i.e., no skill) becoming more negative over Iberia. The implications of this to me are unclear.
- We agree that changes in regions where skill is non-significant, such as over Iberia, are not meaningful in terms of predictive performance, and have removed this part of the sentence. We also acknowledge that the small areas of significant skill over central Germany may be influenced by sampling variability. We have simplified this sentence including only the significant ResCorr values for eastern Europe in this season:
 - Page 12, lines 318-320: “The ResCorr indicates that initialization enhances skill over eastern Europe in the annual mean (Figure 3k), over northern Germany and western France in MAM (Figure 3m), and over eastern Europe in JJA (Figure 3n).”
- Page 12, lines 305-308: “Overall, model initialization affects WCF skill mainly in winter and summer: it leads to a marked reduction of skill in DJF, while it enhances skill in JJA.”
 - I am not fully convinced on the evidence provided to justify this statement. ‘Marked reduction of skill in DJF, as mentioned above, suggests that there was some predictable skill present in the first place, which there is not in Figure 3b. Similarly, the enhancement of skill in JJA only appears to occur over Northern Europe. I would consider rephrasing this statement to be more precise.
- We agree that the original statement was too strong and could be misleading, particularly because it implied a reduction of skill in regions where DCPD skill is not statistically significant. We have therefore removed this statement from the manuscript. Rather than making a broad conclusion about seasonal changes in skill, the revised text focuses on the regional patterns of ResCorr and highlights only the areas where initialization shows a clear and statistically supported impact. This provides a more accurate description of the results and avoids overinterpreting the differences between DCPD and HIST.
- Pg 12-13, lines 321-323: “In addition, ResCorr is significantly positive in Central Europe in MAM, even though the correlation is not significant and trend agreement is weak, indicating the increase of skill due to the initialization in this case.”
 - Could the significance in this case potentially be by chance? As with 95% significance we would expect 1 in 20 samples to show significant correlations by chance. This result is not reflected in the spatial plot in Figure 3h.
- We agree that, although the spatially averaged ResCorr over Central Europe is significantly positive, the signal is spatially limited and not consistently reflected in the corresponding map. This suggests that the regional significance may be influenced by localized areas of positive values and should therefore be interpreted with caution. We have revised the text accordingly to avoid overinterpretation.

- Page 14, lines 336-338: “In addition, ResCorr is significantly positive in Central Europe in MAM; however, this signal is driven by localized positive values (Figure 3h) and may be influenced by sampling variability.”
- Pg 13, lines 323-324: “Other regions show non-significant but positive ResCorr values, suggesting local improvements due to initialization.”
 - I am not clear as to whether insignificant residual correlation values of 0.1-0.2 suggest local improvements due to initialisation. As these are such low values, they could be the result of noise. Additionally, where there is no correlation skill shown in the region in Figure 4c, skill does not improve, rather negative skill becomes less negative. Perhaps it would be more appropriate to focus the discussion on findings where there is both positive skill in Figure 4c and positive residual correlation skill in Figure 4d.
- We agree that non-significant and low positive ResCorr values do not necessarily indicate meaningful improvements in predictive skill. In particular, when correlation skill remains negative, a positive ResCorr may simply reflect a reduction in negative skill rather than a transition to useful predictive performance. We have revised the text to clarify this point and to focus the discussion on cases where both correlation skill and ResCorr indicate consistent improvements.
 - Page 14, lines 338-340: “Other regions show non-significant but positive ResCorr values, suggesting local improvements due to initialization. In regions with weak or negative correlations, however, they mainly reflect a reduction in negative skill rather than a transition to meaningful predictive skill.”
- Pg 13, lines 330-331: “in other cases providing additional skill independently of it.”
 - I am not clear on where initialisation provides additional skill independent of agreement in the trends. Is this referring to the example of MAM Central Europe in Figure 4d?
- We agree that the previous wording was unclear and may have suggested an interpretation that was not sufficiently supported by the results. In the revised manuscript, we removed this statement. We also reduced the emphasis on cases where ResCorr is positive but ACC remains non-significant, since such situations can arise from sampling variability or from a reduction in the magnitude of negative correlations rather than from a clear gain in predictive skill. The revised discussion therefore focuses on regions where both the ACC and ResCorr analyses provide consistent evidence of an added value from initialization.
- Pg 13, lines 332-333: “Using climate projections, Pryor et al. (2005) have identified slight decreases in wind speed and wind energy density over northern Europe.”
 - Are CMIP3 class models a fair comparison for the results shown here? Have there been other, more recent, studies that support the findings of this research?
- We agree that Pryor et al. (2005), based on CMIP3 models, is not directly comparable with the CMIP6-based simulations used in the present study. We have therefore revised this section to include more recent CMIP6-based analyses. In particular, Miao et al. (2023) reported historical decreases in surface wind speed across Europe during 1979–2014, identifying Europe as the Northern Hemisphere region with the strongest declining trends in several CMIP6 models, including the

ones analysed in the present work. We have included this reference in the manuscript.

- Page 14, line 341-343: “Using CMIP6 historical simulations, Miao et al. (2023) identified negative historical wind speed trends over Europe across a broad CMIP6 ensemble, including the three CMIP6 models analysed in the present study.”
- Pg 13-14, lines 336-338: “Observational analyses by Vautard et al. (2010) and Torralba et al. (2017a) have revealed spatially heterogeneous near-surface wind trends across Europe, with predominantly negative signals in several regions.”
 - What level of agreement is there between the observed regional trends shown here and in the studies referenced? Do the areas of trend increase/decrease shown in Figure 4c line up with previous analysis in this area?
- The spatial patterns of wind speed trends shown in Figure 4c are in partial agreement with previous studies. In particular, the negative trends over central and eastern Europe and the weak or heterogeneous signals over southern regions are broadly consistent with both Torralba et al. (2017) and Vautard et al. (2010).

However, some discrepancies are found. In winter, differences are present with respect to Torralba et al. (2017), particularly over northern Europe and the British Isles, where our results show positive trends while their analysis indicates neutral or negative trends. Differences are also present with Vautard et al. (2010) over northern Europe and the British Isles, where negative trends are reported in their ERA-Interim analysis but positive ones are found here. These differences are likely related to the use of different datasets, time periods and methodological approaches. We have clarified this comparison in the revised manuscript.

- Page 14, lines 346-350: “Observational and reanalysis-based studies (e.g. Vautard et al., 2010; Torralba et al., 2017a) have reported spatially heterogeneous near-surface wind trends across Europe. Our results are broadly consistent in showing negative trends over central and eastern Europe and weak signals over southern regions. However, differences remain over northern Europe and the British Isles, where these previous studies reported neutral or negative trends while positive trends are found here, likely due to differences in datasets, periods and methodologies.”
- Pg 14, lines 360-361: “In our case, the prior calibration of all climate variables against ERA5 ensures consistent magnitudes...”
 - Consistent magnitudes of what? Magnitudes of variability?
- We agree that the wording was unclear. We have revised the text to explicitly state that the calibration aligns the statistical distributions of the variables with ERA5, including both mean and variability, which supports the use of absolute thresholds.
 - Page 15, lines 367-369: “In our case, the prior calibration of all climate variables against ERA5 reduces systematic biases relative to the observational reference, including biases in mean values and variability. This helps avoid artificial shifts in the frequency of threshold exceedances caused by biases in the raw model data.”

- Pg 14-15, lines 370-373: “By contrast, both-Neff exhibits more spatially fragmented skill, with isolated regions of significant values. In solar-Neff and both-Neff, some areas appear as missing values because correlations cannot be computed when the indicator remains equal to zero for all years, and solar energy droughts are expectable when the amount of daily sun hours is low (Figure S13b, e).”
 - I am not clear on the presentation of NED for DJF in Figure 5. Figure S13b, g, and i, suggest that for much of Europe, the solar indicator is equal to zero for all years, due to the low amount of daily sunshine hours. In this case, most of the area in Figure 5b and 5g is solely dependent on the number of ineffective days for the wind resource, as the number of ineffective days for solar is constant (i.e., all winter days are ineffective for solar in the greyed out region in Figure S13). Therefore, this would suggest that most of the area in Figure 5b and 5g does not show the combined NED index, as only one of the variables (e.g., wind speed) varies. It would be useful to clarify this and potentially modify the figure to show only the area where the NED index incorporates both solar and wind speed, as this may be misleading to the reader otherwise.
- As noted above, we agree that during DJF the solar component of the NED index shows very limited variability over most of Europe, as solar-Neff is effectively zero in large regions. As a consequence, the variability of NED in this season is primarily driven by the wind component. However, NED remains a compound indicator by definition, as it is based on the joint occurrence of ineffective wind and solar conditions. In winter, the lack of effective solar production implies that this condition is always fulfilled in a big part of the region, so NED effectively reflects the occurrence of ineffective wind days. This represents a seasonal regime of the compound indicator rather than a change in its definition. A similar, though more limited, situation occurs in some northern regions during SON. We have revised the manuscript to clarify this interpretation and avoid potential misinterpretation of the DJF results.
 - Page 15, line 375-386: “Figure S16 presents the corresponding ACC for solar-Neff (Figure S16a–e) and wind-Neff (Figure S16f–j). The skill patterns of solar-Neff closely resemble those of PVpot (Figure 1a–e), while wind-Neff shows patterns similar to those of WCF (Figure 3a–e). Consequently, the spatial distribution of NED skill reflects the seasonally varying contribution of solar- and wind-related conditions. For the annual mean (Figure 5a), significant NED skill over parts of eastern Europe and Iberia broadly coincides with regions where both solar-Neff (Figure S16a) and wind-Neff (Figure S16f) show positive skill. Over parts of southeastern Europe, including the Carpathian region and Greece, localized NED skill is more consistent with solar-Neff than with wind-Neff. In MAM (Figure 5c), NED skill over southern Europe generally follows the solar-Neff signal, although NED remains non-significant in several regions where solar-Neff shows skill. During JJA (Figure 5d), the widespread NED skill closely resembles the solar-Neff pattern, with lower values over parts of western Europe. Over most of Europe in DJF and in northern regions in SON, as solar-Neff remains close to zero (Figure S14b,e), ACC can not be computed (Figure S16b,e) and NED ACC is directly the wind-Neff one (Figure S16g,j). In SON, outside these regions, the

weak NED skill is more consistent with wind-Neff (Figure S16j), although some contribution from solar-Neff (Figure S16e) is also apparent.”

- Pg 15, lines 376-378: “By contrast, over parts of the Alps and southeastern region, NED exhibits significant skill despite wind-Neff being non-significant, indicating a dominant contribution from solar-Neff.”
 - In Figure 5a (annual NED correlations), I cannot clearly see a significant area of positive skill over the Alps. While there is some skill over eastern Europe, both solar-Neff and wind-Neff appear to show some significant positive skill over this region in Figure S13, although the magnitude of the positive skill is lower for wind-Neff.
- We agree that the reference to the Alps is not clearly supported by the figure and has been removed. Regarding the regional description, we acknowledge that the term “southeastern Europe” was too imprecise. We intended to refer to areas spanning eastern and southeastern Europe, including the Carpathian region and Greece, where NED shows localized areas of significant skill. In these regions, the signal is more consistent with solar-Neff than with wind-Neff, which generally shows weaker or non-significant values. We have revised the text to clarify this point.
 - Page 15, lines 379-381: “Over parts of southeastern Europe, including the Carpathian region and Greece, localized NED skill is more consistent with solar-Neff than with wind-Neff.”
- Pg 15: lines 378-379: “In DJF (Figure 5b), NED skill over Iberia and eastern Europe aligns mainly with wind-Neff skill (Figure S13g).”
 - As with the point described above for lines 370-373, I think this statement could be misleading, as only wind-neff skill varies over eastern Europe and much of Iberia. The solar-neff is effectively constant. In this case, surely the NED skill over eastern Europe can only align with wind-neff skill, as solar-neff is constant?
- We agree that, during DJF, NED variability over large parts of Europe is effectively controlled by wind conditions because solar-Neff remains equal to zero in many regions. This point has been clarified earlier in the revised paragraph describing Figure 5 and Figure S16, avoiding a separate interpretation of DJF NED skill solely in terms of wind-Neff correspondence.
- Pg 15, lines 381-383: “Finally, in SON (Figure 5e), NED skill is confined to parts of Poland and the Baltic region, closely matching the wind-Neff skill pattern (Figure S13j) and in points of southeastern Europe, which are skillful at solar-Neff (Figure S13e).”
 - Much of the NED skill shown in Figure 5e is non-significant here. Additionally, both solar-Neff and wind-Neff show skill over Poland, suggesting that both sources contribute, rather than only the wind-Neff mentioned in this statement here. Additionally, the area of significant skill in southeastern Europe is small, suggesting that it could be related to noise.
- We agree that the SON pattern was overinterpreted in the original text. NED skill in SON is weak and mostly non-significant, so any attribution to a single component should be made cautiously. The revised text now states that the SON NED pattern is more consistent with wind-Neff, while acknowledging that solar-Neff may also

contribute in some regions. We also avoid interpreting the small isolated areas over southeastern Europe as robust skill.

- Page 15, lines 372-373: “In SON (Figure 5e), skill is generally weak and confined to parts of eastern Europe and the Baltics.”
- Page 15, lines 385-386: “In SON, outside these regions, the weak NED skill is more consistent with wind-Neff (Figure S16j), although some contribution from solar-Neff (Figure S16e) is also apparent.”
- Page 15, Lines 385-386: “During periods of high solar availability (JJA and, in the southern region, MAM), NED skill is largely driven by solar-Neff, whereas in low-radiation seasons (DJF and SON) wind-Neff becomes the main modulator.”
 - As with the comments above, I think this conclusion may be misleading for the DJF result, as, for much of the European area shown, only the number of ineffective wind days contributes, as the number of ineffective solar days is effectively constant. For much of the region in DJF therefore, wind-(in-)Neff is the only modulator, so it is misleading to suggest that it is the main modulator. This point is not clarified in enough detail in the main text. Additionally, the suggestion that wind-Neff is the main modulator for SON may also be misleading, as this is only true for the limited area of skill over southeastern Europe, while for Poland/Baltic area, both the wind-Neff and solar-Neff appear to contribute.
- We agree that the original statement was overly general given the spatial variability and complexity of the results. Following the reviewer’s suggestions, we consider that a generalised conclusion on the seasonal modulation of NED based on a single dominant energy source is not sufficiently supported. We have therefore removed this paragraph in the revised manuscript to avoid overinterpretation and keep the discussion focused on the regionally consistent and robust features of the results.
- Page 16, lines 393-395: “In MAM (Figure 5h), initialization enhances skill over parts of southern and eastern Europe while reducing it over Iberia, leading to a modest improvement of DCPD skill relative to HIST (Figure 5c vs Figure S14c).”
 - Much of the ResCorr skill in Figure 5h is marginal and insignificant, suggesting that these areas could be due to chance. The difference in significant area of ~4% is relatively small. Therefore, the improvements from initialisation are marginal and the DCPD skill is approximately the same as the HIST (i.e., initialisation does not significantly degrade the skill).
- We agree that, in MAM, the ResCorr values are generally weak and spatially heterogeneous. This suggests that the apparent changes associated with initialization are likely influenced by sampling variability rather than reflecting a robust signal. We also acknowledge that the difference in the fraction of significant area between DCPD and HIST is small, and therefore does not support a clear improvement in skill due to initialization. We have revised the text accordingly to provide a more cautious interpretation.
 - Page 16, lines 391-392: “ResCorr values are generally weak and spatially heterogeneous in the rest of the seasons”
- Page 16: lines 398-400: “For this indicator, initialization generally increases predictability under conditions of higher sunshine duration (JJA and MAM) and reduces it under the opposite conditions (DJF and SON).”

- Once again, any findings related to solar-Neff driving DJF results are misleading, as only the number of inefficient wind days contributes across most of Europe. The only season where initialisation appears to increase skill via ResCorr in Figure 5 is JJA, with MAM, DJF, and SON all showing few areas with any ResCorr skill. I would suggest this sentence is rephrased for clarity, as the MAM ResCorr result is not equivalent to the JJA ResCorr result.
- We agree that the original statement was too general and could be misleading. The clearest and most consistent impact of initialization is found in JJA, where DCPD shows higher skill than HIST. In the other seasons, including MAM, DJF and SON, the ResCorr signal is weak, spatially limited, or not robust, and does not support a consistent interpretation of either improvement or degradation of skill. We have removed the general conclusion about these patterns.
- Page 16, lines 403-404: “In JJA, Iberia, Western Europe, Mediterranean, Central and Eastern Europe display significant negative trends in both datasets, which results in significantly positive correlations (Figure 6c).”
 - It would be useful here to understand the relative contributions of trends in the number of ineffective wind and solar days, in absolute terms. I suspect that the trends in solar radiation contribute more strongly than the trends in wind speed. It would be useful to include greater detail on this here.
- We agree that analysing the relative contributions of the wind and solar components provides important insight into the seasonal behaviour of NED trends. To address this, we have included a new analysis of trends in solar-Neff and wind-Neff (Figure A3, Figure S18 in the manuscript), which allows a more direct physical interpretation of the results. This analysis shows that in JJA the strongest NED trends are primarily associated with changes in solar-Neff, while wind-Neff exhibits weaker and less coherent signals. The manuscript has been revised accordingly.

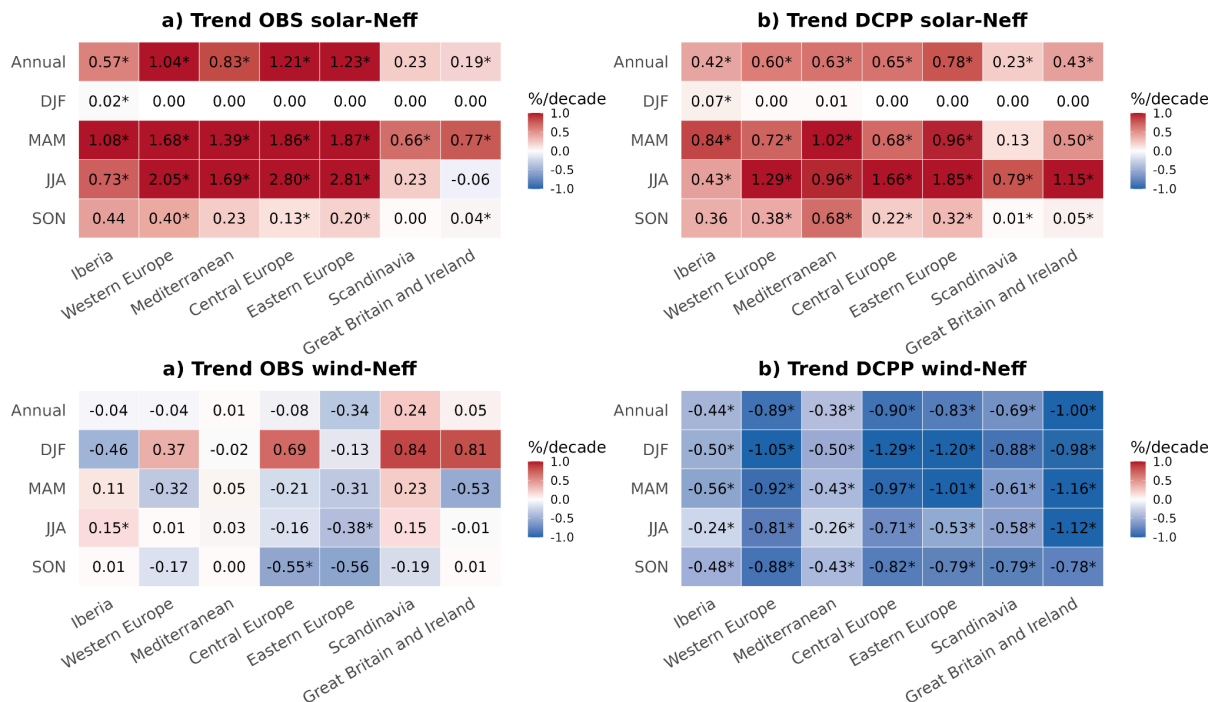


Figure A3. Regional trends for ERA5 (first column) and DCPD (second column) for solar-Neff (a-b) and wind-Neff (c-d). Statistically significant trends using a modified Mann–Kendall test accounting for temporal autocorrelation at the 95% confidence level are marked with an asterisk.

- Page 17, lines 413-421: “To better understand the contribution of solar-Neff and wind-Neff to NED variability and trends, solar-Neff and wind-Neff trends are analysed separately (Figure S18). The results show a clear seasonal dependence. In JJA, the strong negative NED trends in both ERA5 and DCPD are primarily associated with marked increases in solar-Neff (Figure S18a-b), while wind-Neff changes are weaker (Figure S18c-d). In MAM, solar-Neff also drives NED trends in ERA5, although with smaller magnitude, whereas in DCPD this signal is less clear due to strong negative trends in wind-Neff. In DJF, solar-Neff shows negligible variability over most regions, and NED trends are therefore directly anticorrelated with wind-Neff. In SON, both components exhibit mixed signals in ERA5, while in DCPD negative wind-Neff trends tend to dominate, leading to positive NED trends. In the annual means, the strong negative trends in NED in ERA5 seem to be dominated by positive solar-Neff trends, while the mixed DCPD signals are influenced by both energy sources.”
- Page 16: lines 418-419: “NED variability is largely controlled by the energy source that dominates effective production in each season: wind in DJF and SON, and solar in JJA and, secondarily, in MAM.”
 - See the above points on the number of ineffective wind days in DJF being the only contributor to NED here.
- We agree that the original statement was too general, particularly for DJF. In this season, solar-Neff shows very limited variability over most of Europe, so NED variability is directly anticorrelated with wind conditions. Following the revision described above, the interpretation of NED variability has been reformulated based on the direct analysis of solar-Neff and wind-Neff trends (Figure S18), providing a more precise and seasonally consistent description.
- Figures 2, 4, 6
 - It would be useful to have subplot headings here, to convey the information more clearly to the reader, these could be: “a) ERA5 trends”, “b) DCPD trends”, “c) Corr”, “d) ResCorr”.
- We have added the subtitles to the corresponding figures.
- Page 17, lines 429-430: “Additionally, DCPD trends show seasonal patterns that broadly resemble the future evolution of hybrid energy droughts projected by (Kapica et al., 2024), particularly in DJF and SON.”
 - It would be useful to have further clarification on how these resemble the results of Kapica et al., 2024, particularly regarding the similarities and differences of the NED metric used here compared to the hybrid metric from Kapica.
- We agree that the comparison with Kapica et al. (2024) required further clarification. A better qualitative agreement is found in MAM, where both studies show a north-south contrast (decreasing drought conditions in southern Europe and

increasing conditions in northern regions), and to a lesser extent in JJA. In DJF, some regional agreement is also found, particularly over northern and eastern Europe, where both studies indicate increasing energy drought conditions. In addition, discrepancies remain in southern regions such as Iberia and France, where the two studies show opposite trends. In SON, the agreement is weak, with Kapica et al. (2024) indicating decreasing drought conditions while our results show a general increase.

We have revised the manuscript accordingly and clarify that Kapica et al. (2024) use a percentile-based definition of droughts, whereas the NED metric employed here is based on fixed, physically motivated thresholds. The comparison is therefore limited to qualitative spatial patterns.

The following sentences have been added:

- Page 18, line 424-429: “Additionally, DCPD trends show some qualitative similarities with the future evolution of energy droughts projected by Kapica et al. (2024), particularly in MAM and over parts of northern and eastern Europe in DJF. However, Kapica et al. (2024) have used a percentile-based definition of droughts, whereas the NED metric employed here is based on fixed physically motivated thresholds. The comparison is therefore limited to broad spatial patterns, and important regional differences remain, particularly in SON.”
- Page 18, lines 445-446: “This strong seasonal dependence indicates that PVpot predictability is highest during periods of higher and persistent solar radiation, while such predictability is more limited under weaker and more variable radiative conditions.”
 - To what extent is the highest predictability during the summer driven by the fact that the trends are strongest during the summer (Figure 2)?
- We agree that the higher predictability observed in summer may be highly related to the stronger and more spatially coherent trends shown in Figure 2, which can enhance correlation skill. However, the original statement linked this behaviour to more persistent solar radiation, which was not explicitly assessed in this study. To avoid introducing an unsupported interpretation, we have removed this sentence in the revised manuscript.
- Page 19: lines 455-457: “Significant skill is consistently found over parts of eastern Europe across seasons, while other regions show skill only seasonally: central Europe in JJA, southern Europe in DJF, and northern Europe in MAM and SON.”
 - I am not convinced that significant skill is found over southern Europe in DJF (Figure 3b). While there is some skill over eastern Europe, the remaining areas of significant positive skill are small and spatially heterogeneous.
- We agree that the areas of significant skill over southern Europe in DJF are limited, spatially fragmented, and not particularly robust. We have revised the text to reflect this more accurately.
 - Page 18, line 449-451: “Significant skill is consistently found over parts of eastern Europe across seasons, while other regions show weaker and more

seasonally dependent skill, including central Europe in JJA and northern Europe in MAM and SON.”

- Page 18, lines 459-460: “Initialization tends to enhance skill in summer and reduce it in winter, while its impact is generally small in the transition seasons.”
 - Most of the ResCorr in Figure 3i is over Eastern Europe, suggesting that the enhancement of skill in summer is more regionally constrained than suggested in this statement. The reduction of skill during the winter is mostly negative skill becoming more negative, suggesting that there wasn’t skill to reduce in the first place. Consider rephrasing.
- We agree that the original statement was too general. In JJA, the enhancement of skill is mainly confined to eastern Europe, while in DJF the negative ResCorr values largely reflect already low or negative skill becoming more negative rather than a meaningful reduction of predictive skill. We have revised the text accordingly.
 - Page 19, line 453-454: “The impact of initialization is regionally dependent and generally limited. The clearest skill enhancements are found in JJA over parts of eastern Europe, while in other seasons the impact of initialization is weaker and spatially heterogeneous.”
- Page 19, lines 466-469: “Seasonal differences in NED skill reflect the dominant energy source. During high-radiation periods (JJA and partly MAM), NED predictability is mainly driven by solar-Neff, whereas during low-radiation seasons (DJF and SON) wind-Neff becomes the main contributor.”
 - See the above points on the number of inefficient wind days being the only contributing factor to skill during DJF. Consider rephrasing and/or changing narrative above.
- We agree that the original statement was too general, particularly for DJF. We have revised the text to provide a more precise and regionally consistent interpretation of the relative contribution of wind and solar components across seasons.
 - Page 19, line 463-468: “Seasonal differences in NED skill reflect the relative contribution of wind and solar components. In JJA, NED predictability is mainly driven by solar-Neff. In DJF, variability over much of Europe is largely driven by wind conditions, while in MAM and SON the relative contribution of both components varies regionally. As a result, the impact of initialization on NED is also seasonally dependent, enhancing skill when the dominant energy source is more predictable and reducing it otherwise. This is consistent with the broader finding that initialization generally enhances existing predictability rather than creating skill where historical simulations show none.”
- Page 19, line 480: “However, ERA5 has been identified as the best reanalysis for wind (Ramon et al., 2019)”
 - It would be useful for the reader to be more specific about the particular characteristics of ERA5 which make it preferable to other reanalysis products here.
- We agree that the original statement was too general. Ramon et al. (2019) show that ERA5 outperforms other reanalyses primarily in terms of correlation with observations and representation of temporal variability, with statistically significant improvements over a multi-reanalysis mean in a substantial fraction of locations.

However, differences among reanalyses remain large for mean wind and long-term trends. We have revised the text to clarify these aspects.

- Page 19, line 478-480: “However, ERA5 has been identified as one of the most reliable reanalysis datasets for near-surface wind variability, particularly in terms of correlation with observations and representation of temporal variability at turbine-relevant heights (Ramon et al., 2019)”
- Page 19, lines 492-494: “In addition, hybrid approaches that combine percentile-based metrics with physically motivated thresholds may help balance climatological comparability and operational relevance when assessing compound renewable-energy risks.”
 - How can percentile-based metrics be combined with physical thresholds in a hybrid approach? Does this involve selecting the percentiles based on physical thresholds or vice versa?
- We agree that the original statement was unclear. To avoid introducing ambiguity, we have simplified this part of the discussion.

Technical corrections

- Pg 8 Line 216: “Figure 2 summarizes the regional behaviour of PVpot long-term trends across the seven European sub-regions (Figure S2)”
 - Figure S2 should be Figure S4
- Pg 12, line 301: “and the DCPD (Figure 3c) has higher skill values than HIST (Figure S10h)”
 - Figure S10h should be Figure S10c
- Pg 14, line 353: “...almost all seasons closely resembles the patterns found for the WCF standard deviation (Figure S8f-j),”
 - Figure S8f-j should be Figure S9f-j
- Pg 15, line 373: “and solar energy droughts are expectable when...”
 - Replace ‘expectable’ with ‘to be expected’ or ‘expected’
- All these technical corrections have been changed in the revised version of the manuscript

Chen, W. B. (2025). Assessing global land-based solar–wind complementarity using high-resolution climate reanalysis for hybrid renewable energy design. *Energy Conversion and Management*, 343, 120267.

Dubus, L., Saint-Drenan, Y. M., Troccoli, A., De Felice, M., Moreau, Y., Ho-Tran, L., ... & Sanger, L. (2023). C3S Energy: A climate service for the provision of power supply and demand indicators for Europe based on the ERA5 reanalysis and ENTSO-E data. *Meteorological Applications*, 30(5), e2145.

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.

Hewitson, B. C., Daron, J., Crane, R. G., Zermoglio, M. F., & Jack, C. (2014). Interrogating empirical-statistical downscaling. *Climatic change*, 122(4), 539-554.

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.

Maraun, D., Wetterhall, F., Ireson, A. M., Chandler, R. E., Kendon, E. J., Widmann, M., ... & Thiele-Eich, I. (2010). Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. *Reviews of geophysics*, 48(3).

Maraun, D., & Widmann, M. (2018). *Statistical downscaling and bias correction for climate research*. Cambridge University Press.

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.

Peña, A., & Gryning, S. E. (2008). Charnock's roughness length model and non-dimensional wind profiles over the sea. *Boundary-layer meteorology*, 128(2), 191-203.

Roberts, C. D., & Leutbecher, M. (2025). Unbiased calculation, evaluation, and calibration of ensemble forecast anomalies. *Quarterly Journal of the Royal Meteorological Society*, 151(771), e4993.

Soares, P. M., Cardoso, R. M., Lima, D. C., & Miranda, P. M. (2017). Future precipitation in Portugal: high-resolution projections using WRF model and EURO-CORDEX multi-model ensembles. *Climate Dynamics*, 49(7), 2503-2530.

Staffell, I., & Pfenninger, S. (2016). Using bias-corrected reanalysis to simulate current and future wind power output. *Energy*, 114, 1224-1239.

WMO (World Meteorological Organization). (2011). Climate Knowledge for Action: A Global Framework for Climate Services—Empowering the Most Vulnerable, Report of the High-Level Taskforce for the Global Framework for Climate Services.

Moreno-Montes et al 2026 Reviewer #2 Comments

Title: Decadal predictions of wind, solar and compound power indicators to support the European renewable energy sector

Author(s): Sara Moreno-Montes et al.

MS No.: egusphere-2026-1205

MS type: Research article

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 DCPD and HIST simulations may be contributing to the enhanced DCPD 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.

We would like to thank the reviewer for the careful review of our manuscript and for the valuable comments and suggestions. These have helped us improve the clarity of the manuscript, strengthen several methodological aspects, and refine the discussion of the results. Our detailed point-by-point responses to each comment are provided below (in blue).

General comments:

1. As mentioned above, my main concern is that the interpretation of differences in skill between DCPD and HIST simulations is due to initialisation could be hampered by the different ensemble sizes: The DCPD 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 DCPD 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 DCPD may influence the comparison with HIST by reducing internal variability and improving the representation of the forced signal. We have therefore added an additional sensitivity test to assess whether the diagnosed added value of DCPD over HIST is affected by the different ensemble sizes.

Our main results are based on the multi-model mean. For each model, we first average across ensemble members, and then average across models, so that each model contributes equally to the final multi-model estimate. For this reason, the resampling was also performed separately for each model rather than by randomly sampling members from

the full pooled ensemble. This preserves the weighting structure used in the main analysis and avoids giving disproportionate weight to models with more available members.

For each model family, we randomly sampled the same number of DCPD and HIST members, using the minimum number of available members in the two systems. Specifically, after pooling the three EC-Earth3 DCPD initializations, the available ensemble sizes were 24 DCPD members for EC-Earth3, 10 for IPSL-CM6A-LR, and 10 for MPI-ESM1.2-HR. The corresponding HIST ensembles contained 7, 25, and 2 members, respectively. We therefore used 7 EC-Earth members, 10 IPSL members, and 2 MPI members in both DCPD and HIST. The selected members were then averaged within each model, and the three model means were averaged to construct the multi-model mean, exactly as in the main analysis.

This procedure was repeated 5000 times. For each resampling iteration, we recalculated the residual correlation between DCPD and ERA5 relative to HIST. We then compared this equal-member residual correlation with the residual correlation obtained using the full ensembles:

$$\Delta ResCorr = ResCorr_{equal-members} - ResCorr_{full-ensemble}$$

This quantity directly measures the sensitivity of the diagnosed DCPD added value to the unequal ensemble sizes. Values close to zero indicate that the ResCorr results are not substantially affected by ensemble-size differences. Negative values indicate that reducing the DCPD ensemble size decreases the added value of initialization, whereas positive values indicate the opposite.

For each grid point, season and indicator, we summarized the 5000 resampling estimates using the median $\Delta ResCorr$. Additionally, we also computed the 2.5th and 97.5th percentiles of the resampling distribution to test the statistical significance. Regions where this 95% bootstrap interval did not include zero were identified as regions where the ResCorr estimate is significantly sensitive to the ensemble-size adjustment. These new analyses have been added at the Supplementary material in Figures S8, S12 and S17.

Page 6, lines 176-186: “Since the DCPD and HIST ensembles have different numbers of members, we perform an additional sensitivity analysis to assess the impact of ensemble size on the estimated added value of initialization. The analysis is applied to the multi-model mean used throughout the study; therefore, ensemble-size adjustments are performed independently for each model before constructing the multi-model average. For each model family, the same number of DCPD and HIST members is selected, corresponding to the minimum ensemble size available in both systems (7 for EC-Earth3, 10 for IPSL-CM6A-LR, and 2 for MPI-ESM1.2-HR). For this analysis, the 7 EC-Earth3 decadal prediction members are taken from the three different initializations for their comparison with the non-initialized EC-Earth3 historical simulations. For each resampling, an equal-member multi-model ensemble is constructed and used to estimate ResCorr. The procedure is repeated 5000 times, and the sensitivity to ensemble size is quantified as the difference in residual correlation:

$$\Delta ResCorr = ResCorr_{equal-members} - ResCorr_{full-ensemble}$$

The median of the resulting $\Delta ResCorr$ distribution is used as the central estimate at each grid point, while the 2.5th and 97.5th percentiles define a 95% bootstrap interval. Regions where this interval does not include zero are considered robustly sensitive to ensemble size.”

Page 9, lines 249-251: “Figure S8 shows that the median ResCorr differences obtained when subsampling the ensembles are generally close to zero and non-significant across Europe. This indicates that the added value of initialization identified in Figure 1k–o is largely insensitive to the unequal ensemble sizes of the prediction systems.”

Page 13, lines 322-323: “Figure S12 further shows that the ResCorr patterns are largely insensitive to ensemble size, with generally small and non-significant differences across Europe.”

Page 15, lines 396-398: “Figure S17 shows that the ResCorr differences obtained when subsampling the ensembles are generally close to zero and non-significant across Europe, indicating that the NED results are largely insensitive to the unequal ensemble sizes of the prediction systems.”

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 have revised the text to focus on the main messages, removing or simplifying sentences that describe weak or non-significant patterns, including the example highlighted.

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 has been added to the manuscript:

- Page 2, lines 48-50: “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 DCP. 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 DCP.

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

- Page 3, lines 69-70: “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 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 generate additional large-scale predictive information. Nevertheless, the objective of this study is to provide information at the spatial scale required by energy applications, which motivates the use of the downscaled products throughout the main analysis.

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), which showed that more complex downscaling techniques do not systematically provide additional predictive skill compared to simpler approaches.

To assess the sensitivity of the results to spatial resolution, we additionally repeated the analyses on the native resolution of the coarsest prediction system. The corresponding results are presented in Figures S6, S11 and S16 and are explicitly discussed throughout the Results section, where the main skill patterns identified in the downscaled analysis are compared with those obtained at native resolution.

- Page 3, lines 88-95: “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. To assess the sensitivity of the results to the spatial resolution used for the analysis, all indicators and skill metrics are additionally computed on the native spatial resolution of the coarsest prediction system considered (IPSL-CM6A-LR: $2.5^{\circ} \times 1.25^{\circ}$)."

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 have clarified this in the manuscript:

- Page 4, line 101-103: "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 have clarified this in the manuscript.

- Page 4, lines 110-119: "The adjusted decadal predictions and historical simulations are then calibrated against ERA5 using EQM (Panofsky and Brier, 1958). By

adjusting empirical quantiles at 1% intervals (from the 1st to the 99th percentile), EQM corrects the distribution of the variable, including the mean, variance, and higher-order moments. Calibration is performed independently at each grid point, per forecast year and member for decadal predictions, and per calendar year and member for historical simulations. In both cases, empirical distributions are estimated using a 30-day moving window centered on the target date to preserve seasonality while ensuring a sufficient sample size. A leave-one-year-out cross-validation is applied. For each target year, the climatological adjustment and EQM transfer function are estimated using all other years in the calibration period. This prevents information from the target year from being used in the calibration and avoids artificial skill inflation (Elsner and Schmertmann, 1994). For 6-hWIND, the calibration is performed separately for each time of the day, since wind distributions differ markedly across the diurnal cycle (Kalverla et al., 2019; Weide Luiz and Fiedler, 2022).

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 DCPD and HIST skill is not straightforward when the information is split between the main figures and the supplementary material. To improve clarity, we have revised the figures to facilitate a more direct comparison, adding the HIST ACC results in the main panels. This allows a clearer assessment of the differences between DCPD 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 have been included in all the figures.

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 have revised 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 DCPD of the indicator and ERA5" - do you mean "the ACC of the indicator between DCPD and ERA5"? Similar phrases appear elsewhere, e.g. L174.

We agree that the phrasing was unclear. We have revised the text to improve clarity and ensure that it consistently refers to the ACC of the indicator computed between DCPD and ERA5.

- Page 7, lines 201-203: “The ACC of the indicator between DCPD and ERA5, the ACC between HIST and ERA5, and the ResCorr between DCPD and ERA5 relative to HIST are shown for the annual mean and individual seasons.”
- Page 7, lines 203-205: “Additionally, the ACC between DCPD and ERA5 and the regional trends of the variables used to calculate the indicators (RSDS, TAS and SFCWIND) are shown in Figures S7 and S9.”
- Page 8, caption Figure 1: “ACC between DCPD and ERA5 for PVpot for the annual mean...”
- Page 15, lines 370: “Figure 5a–e shows the ACC between DCPD 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 have revised the manuscript accordingly to avoid confusion.

- Page 7, lines 210-212: “The associated multiannual standard deviation is low across the domain (S5f–j), indicating that PVpot variability is small and largely controlled by mean climatological conditions.”
- Page 11, lines 285-287: “Figure S10 shows the climatology (Figure S10a–e) and the multiannual standard deviation (Figure S10f–j) of WCF for the annual and seasonal means.”
- Page 11, lines 288-290: “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 355: “Figure S13 shows the climatology and multiannual standard deviation of NED”

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

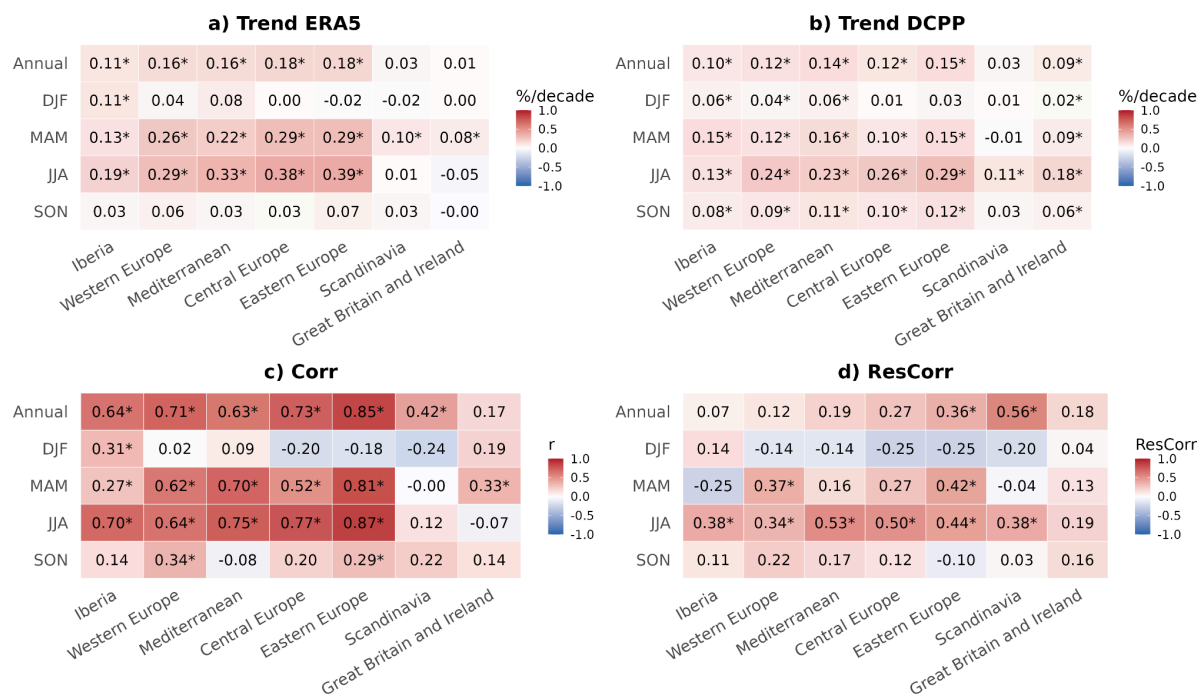
We agree that the panel labelling in Figure S5 (now S7) was wrong. We have revised the figure to ensure that the ordering and labelling of the panels are consistent and that the variables (TAS and RSDS) are clearly identified.

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 have updated 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, lines 193-195: “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).”

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. We have revised the text to clarify the results.

- Page 9, lines 255-257: “In contrast, correlations are mostly weak and non-significant in DJF and SON, when observed trends are small or absent. Scandinavia and Great

Britain and Ireland exhibit little skill overall, with significant correlations only in a few cases.”

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

We thank the reviewer for pointing this out. The figure caption has been changed accordingly.

8. L438 “multi-model ensemble”

“Multi-model” should read “multi-model ensemble”. We have corrected 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-Matamoras, R., & Soret, A. (2026). Comparative analysis of statistical downscaling methods for multi-model decadal climate predictions over Western Europe. *Climate Services*, 42, 100639.