

Reviewer #1

Specific Comments

- Page 2, lines 26-27: “the intermediate horizon of next years remains comparatively underexploited”
 - Surely this should be “the intermediate horizon of the next few years remains comparatively underexploited”?

Thank you for pointing this out. The text has been corrected in the manuscript

- Page 6, line 179-180: “For each forecast system, the same number of DCPD and HIST members is selected, corresponding to the minimum ensemble size”
 - “...the same number of DCPD and HIST members are selected, corresponding...”

The text has been corrected in the manuscript

- Page 9 lines 246-247: “During MAM and SON (Figure 1m,o), positive ResCorr values are restricted to specific regions, mainly over eastern Europe and parts of the Atlantic coast in MAM and over northern Atlantic regions in SON”
 - I am not clear where this ‘northern Atlantic’ region of rescorr skill is visible in Figure 1. Where is this region?

Thank you for pointing this out. The term “northern Atlantic regions” was imprecise. We were referring to the positive ResCorr values around Belgium and the Netherlands. We have revised the text.

Page 9 lines 261-262: “During MAM and SON (Figure 1m,o), positive ResCorr values are restricted to specific regions, mainly over eastern Europe and parts of the Atlantic coast in MAM and over parts of northwestern Europe in SON.”

- Page 11, lines 285-291: “Figure S10 shows the climatology (Figure S10a–e) multiannual standard deviation (Figure S10f–j) of WCF for the annual and seasonal means. The climatological patterns follow a distribution of wind resources over Europe, with higher WCF values over the northern Europe and lower values over southern and Mediterranean regions, together with a clear seasonal cycle (Bett and Thornton, 2016) and higher values offshore than onshore. Unlike PV_{pot}, 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 year-to-year wind variability. This reflects the intrinsically high variability of wind and the non-linear turbine power curve, which translates wind fluctuations into larger variations in WCF.”
 - This is more of a structuring point than a science point, however I think the results section would flow better if you were able to discuss the findings shown in the main figures first (e.g., Figures 1, 3, and 5) and then use evidence from the supplementary figures to support this, rather than introduce the climatological mean and standard deviation (from the supplementary information) before then discussing the main figure. I appreciate this may not be possible in all parts of the paper, however I do think that restructuring in this way would make it easier for the reader.

The subsections of the Results section have been reorganized to present the skill of each indicator first, followed by its climatology and multiannual standard deviation, as suggested.

- Page 11, lines 299-300: “The spatial distribution and the percentage of significant grid points for WCF closely resemble those obtained for SFCWIND (Figure S9k-o)”
 - Looking at the supplementary material, I believe this should refer to Figure S7k-o, rather than S9k-o.

The text has been changed in the manuscript.

- Page 13, lines 318-319: “HIST exhibits higher skill than DCPD in DJF (Figure 3g), particularly over Mediterranean regions. In contrast, HIST shows slightly lower skill than DCPD in MAM and JJA (Figure 3h-i)”
 - I accept that the skill over Iberia and Northern Africa appears stronger for the HIST result in 3g compared to the DCPD result in 3b, however in the MAM and JJA case I think the magnitude of skill (i.e., the red shading) is being conflated with the area of significant positive skill. Perhaps it would be more useful to be more specific about whether you are referring to the significant area % or the magnitude of the skill here.

Thank you for pointing this out. We were referring to the percentage of significantly positive grid points rather than the magnitude of the ACC. We have revised the text to make this distinction explicit.

Page 13, lines 338-340: “HIST exhibits a higher percentage of significantly positive grid points than DCPD in DJF (Figure 3g), particularly over Mediterranean regions. In contrast, the percentage of significantly positive grid points is slightly lower for HIST than for DCPD in MAM and JJA (Figure 3h-i).”

- Page 15, lines 388-390: “Overall, HIST reproduces the main spatial pattern found in DCPD, although the relative skill varies by season. HIST shows slightly lower skill than DCPD in MAM (Figure 5h) and particularly in JJA (Figure 5i), whereas it has higher skill in DJF (Figure 5g)”
 - As mentioned above, I feel that the magnitude of skill is being conflated with the area of significant positive skill. To my eye, generally the magnitudes of skill look similar (i.e., similar shades of red) between the DCPD and hist in Figure 5 for DJF, MAM, and JJA. However, the fractions of significant area are different. Consider rephrasing for greater clarity on what exactly you are referring to with ‘higher’ or ‘lower’ skill here.

Thank you for pointing this out. We agree that the previous wording could conflate the magnitude of skill with the fraction of significantly positive grid points. We have rephrased the sentence to clarify that the comparison refers to the latter.

Page 16, lines 411-413: “HIST shows a slightly lower percentage of significantly positive grid points than DCPD in MAM (Figure 5h) and particularly in JJA (Figure 5i), whereas this percentage is higher for HIST in DJF (Figure 5g).”

- Page 18, lines 414-416: “In JJA, the strong negative NED trends in both ERA5 and DCPD are primarily associated with marked increases in solar-Neb (Figure S18a-b), while wind-Neb changes are weaker (Figure S18c-d)”

- Currently Figure S18 is labelled a) b) a) b), please could you update the bottom row to c) d) for clarity here.

The panel labels have been updated in the manuscript.

Reviewer #2

I appreciate the thorough response the authors have provided to my initial review and I think the manuscript is much improved. I think the experiment the authors have performed to investigate the impact of the larger ensemble size for DCPD compared to HIST is good and I'm satisfied that they've demonstrated it only produces a small effect on the value of ResCorr. However, on a related note, it has recently come to my attention that the paper cited for describing the ResCorr method, Smith et al (2019), notes that the finite size of the HIST ensemble (rather than the relative size between the two ensembles) introduces a positive bias in ResCorr. Smith et al (2019) describe a method to produce an unbiased estimate in ResCorr – please could the authors confirm whether their figures provide the biased or unbiased estimate? I apologise for only bringing this up in a second review, but I think it could be important for the scientific integrity of the paper. It would also probably help the readers if some brief information on how ResCorr is calculated is given.

Thank you for raising this point. The ResCorr values shown in the original manuscript corresponded to the standard estimate obtained by removing the variability associated with HIST from both DCPD and ERA5 and correlating the resulting residuals. We agree that, as discussed by Smith et al. (2019), the finite size of the HIST ensemble can introduce a positive bias in ResCorr. We have therefore added a brief description of the ResCorr calculation and significance testing in Methods and performed an additional analysis to assess this finite-ensemble bias. The main figures continue to show the standard ResCorr estimate, while the bias-corrected estimates and their associated uncertainty are presented as an additional sensitivity analysis in Figures S8, S12 and S17.

Some adaptation of the Smith et al. (2019) approach was required for our multi-model framework. Since the multi-model mean is constructed by first averaging each prediction system independently and then averaging the resulting model mean estimates, the HIST ensemble of each model is divided independently into two equal subsets (3+3 members for EC-Earth3, 5+5 for IPSL-CM6A-LR, and 1+1 for MPI-ESM1.2-HR). For EC-Earth3 and IPSL-CM6A-LR, which have 7 and 11 HIST members, respectively, one member is randomly excluded in each repetition. The corresponding subsets are then averaged across prediction systems to construct two independent multi-model HIST estimates (A and B), thus preserving the same model weighting used throughout the study.

For each repetition, a biased ResCorr estimate is calculated using HIST A as the reference for both DCPD and ERA5. An unbiased estimate is then calculated using the two independent HIST estimates, with HIST A used as the reference for DCPD and HIST B for ERA5. The finite-ensemble bias is estimated as the difference between these two ResCorr values. The procedure is repeated 5000 times, and the median estimated bias is subtracted from the original full-ensemble ResCorr. The standard deviation of the bias estimates across the 5000 repetitions is used to characterize the uncertainty associated with the ensemble partitioning.

An important limitation is the relatively small number of HIST members available for constructing the independent estimates, particularly for MPI-ESM1.2-HR, where each subset necessarily contains only one member. The estimated correction consequently shows substantial uncertainty in some regions and seasons. We therefore present the bias-corrected ResCorr as an additional sensitivity analysis and explicitly discuss this limitation in the revised manuscript.

Page 6, lines 173-175: “ResCorr is calculated by linearly removing the variability associated with HIST from both DCPD and ERA5 and correlating the resulting residuals. Its significance is assessed using a two-sided t-test with effective degrees of freedom to account for temporal autocorrelation.”

Page 6, lines 191-201: “Additionally, we assess the positive bias in ResCorr associated with the finite size of the HIST ensembles, following Smith et al. (2019). Since the multi-model mean is constructed by averaging each prediction system independently, the HIST ensemble of each model is randomly divided into two equal subsets (3+3 members for EC-Earth3, 5+5 for IPSL-CM6A-LR, and 1+1 for MPI-ESM1.2-HR). For EC-Earth3 and IPSL-CM6A-LR, which have 7 and 11 HIST members, respectively, one member is randomly excluded in each repetition. The corresponding subsets are then averaged across prediction systems to construct two independent multi-model HIST estimates (A and B). For each repetition, a biased ResCorr estimate is calculated using HIST A as the reference for both DCPD and ERA5. An unbiased ResCorr estimate is then calculated using the two independent HIST estimates, with HIST A used as the reference for DCPD and HIST B for ERA5. The finite-ensemble bias is estimated as the difference between these two ResCorr values. The procedure is repeated 5000 times, and the median estimated bias is subtracted from the original ResCorr. The uncertainty associated with the ensemble partitioning is quantified as the standard deviation of the bias estimates across the 5000 repetitions.”

Page 7, lines 219-222: “Finally, the robustness of the ResCorr results is further assessed by examining both the sensitivity to differences in ensemble size between DCPD and HIST and the positive bias associated with the finite size of the HIST ensembles, following the procedures described in Section 3. The corresponding results, including the sensitivity to equal-member resampling, the bias-corrected

ResCorr, and the uncertainty associated with the finite-ensemble correction, are presented in Figures S8, S12 and S17.”

Page 10, lines 266-272: “However, accounting for the positive bias associated with the finite size of the HIST ensembles substantially reduces ResCorr, particularly for the annual mean, MAM and JJA (Figure S8f–j), indicating that the standard ResCorr may overestimate the added value of initialization. Nevertheless, the bias-corrected ResCorr is also subject to considerable uncertainty. The correction requires splitting the already small HIST ensembles into two independent subsets, including only one member per subset (1+1) for MPI-ESM1.2-HR. The relatively large standard deviation (Figure S8k–o), particularly in MAM and JJA, reflects this uncertainty. Therefore, both the standard and bias-corrected ResCorr estimates should be interpreted with caution.”

Page 14, lines 345-350: “Accounting for the finite size of the HIST ensembles modifies ResCorr locally but generally preserves its weak and spatially heterogeneous character (Figure S12f–j).”

Page 16, lines 422-424: “Accounting for the finite size of the HIST ensembles substantially reduces some of the strongest ResCorr patterns, particularly for JJA (Figure S17i). The uncertainty of this correction is also relatively high (Figure S17n), indicating that the magnitude of the finite-ensemble correction should be interpreted with caution.”

Page 20, lines 503-507: “Accounting for this finite-ensemble bias generally reduces positive ResCorr values, indicating that the standard estimate may overestimate the added value of initialization in some regions. Nevertheless, the bias correction itself is uncertain due to the small number of members available in each ensemble partition, as reflected by the relatively large standard deviation across partitions. Therefore, the standard and bias-corrected ResCorr estimates should be interpreted with caution.”