

Response to Reviewer 1

Manuscript: “Isolating flow-dependent uncertainty in ensemble reanalysis data and its relation to Euro-Atlantic weather regimes and warm conveyor belts”

We agree with the reviewer that the manuscript ended up being quite lengthy. We thank the reviewer for their detailed and constructive comments and for providing some help to identify parts that may be shortened or even cut in order to improve readability and flow. Responses to each of the comments raised are listed below.

1 Major General Comments

Reviewer comment	Response	Changes in manuscript
1. I found the Methods section lacked some motivation and was unbalanced in the sense that some aspects are discussed in great detail (e.g. error covariance matrices), while there is little discussion of what the covariates might be other than those describing the annual cycle, and the discussion of the breakpoints left me hoping we would be told more later. This meant that I took a lot of time to get through this section. For the residual covariances, I wonder if a couple of images displaying the matrix structures could help? I am not an expert here but, for the GLS approach, might this be a series of diagonal lines of reducing thickness the further away they are from the main diagonal(?). For the HAC approach, perhaps the contours would be deformed into a diagonally-oriented “Y” shape to indicate heteroskedasticity(?)	We agree that the methods section left room for improvement. We have tightened it where necessary, added references to its application where possible and have revised its structure. We also added an introductory paragraph for additional reader guidance. Some details previously in the “Results” section were moved here for consistency. Focussing on what is essential, we have moved the “Model evaluation and inference” section to the appendix. We have also made our consideration about error covariances more concise and clear. We considered introducing the matrix algebra of the GLS fitting procedure to illustrate the implications of the variance and autocorrelation assumptions, but have eventually not been able to introduce it in such a way that would justify opening another tangent—especially so because we have made our consideration of the error covariance structure more brief. We have replaced the term “levels” to “categories” when talking about categorical covariates to avoid confusion. We have also elaborated on how categorical covariates are implemented and evaluated.	Whole “Methods” section.

Reviewer comment	Response	Changes in manuscript
It is only in the Results section (Eq. 19) where we find that the initial GLS is simply aiming to find a linear trend in the log of the EDA standard deviation, while fitting to a 3-harmonic annual cycle. A motivating statement in the Methods section might be something like “We use the generalised least squares approach to take account of seasonality while fitting a linear trend to the data, and we later partition this by the weather regimes”. To make the Methods section accessible to more readers (including me), please discuss briefly how categorical covariates can be incorporated into GLS. The text (later, on lines 674–675) “Figure 7 shows the estimated marginal mean effect (cf. Eq. 13) of each of the WRs, from a model with the WR time series as the sole covariate” seems misleading, in the sense that I suspect the software package used will have replaced this “sole covariate” with dummy binary variables for all (but one) of these categories(?) The description of the marginal means calculation talks about “levels” which, in the readers’ mind, will have a very different connotation to (non-numerical) “regimes”.		

Reviewer comment	Response	Changes in manuscript
2. In relation to Figure 1, the 95 % confidence intervals about the change points seem very narrow. Do these confidence intervals make sense in the light on the constraints imposed on the length and number of segments (within the optimisation of the Bayesian Information Criterion, BIC)? Figure D1 suggests that five segments are also near-optimal based on BIC; how does the timing of the break points change with five segments? In the first segment, the log variance decreases sharply; does this segment exist by virtue of consistent fitting, or is it due to the “minimum segment length” criterion? (The fit does not look brilliant in places). Perhaps, in the light of the variance trends in each segment, the precise locations of the break points are not so important anyway, but rather it is good to account for some discontinuities in the record?	The confidence intervals around each breakpoint are estimated by holding all other breakpoints in place and calculating the range in which the breakpoint lies with 95 % probability according to the limit distribution (with HAC covariance estimators). They do not account for the uncertainty in the number of breakpoints, which is why we add Fig. E1 (formerly D1) in the first place. In a setting with four breakpoints, the last breakpoint (2008-09) disappears with the other ones staying almost exactly at the same point with almost the same CIs. This also points to the question raised about the first segment. We have added a note about the first segment looking peculiar. Since the manuscript is quite lengthy already, we did not add the figure for four breakpoints, but provide it in this response letter as Fig. 1	The fit to the data in this segment is worse than in the other segments, but the second-stage models consider only data from 1950 or 1979 onwards respectively, so the first segment should not confound the insights gained there (see Sect. 5.3).
3. In relation to Figure 7, the authors highlight (and investigate) the negative correspondence between the weather regime anomalies and the spread residuals. Looking closely, the positive residuals actually appear to me to be shifted eastward somewhat relative to the troughs in the weather regimes. See, in particular: AT, ZO, and EuBL. From the literature, it is found that the largest growth in spread tends to be located along the leading edges of troughs (including warm conveyor belts), where moist processes are particularly active. Do the authors think that this might be a plausible explanation of the displayed results? (The negative spread residuals in the anticyclonic anomalies would also follow ipso facto).	We think this aspect raised by the reviewer is convincing and compatible with our understanding of the mechanisms and was, in fact, lacking from the discussion of Fig. 7. We have accordingly added this consideration and thank the reviewer for the valuable addition.	Moreover, positive anomalies seem to be more pronounced at the leading edges of troughs (AT, ZO, EuBL), which are known for intense baroclinic development and the occurrence of WCBs.

Reviewer comment	Response	Changes in manuscript
4. In relation to section 5.3.1 and Figure 8. The authors suggest that poor correlations in the early record between variance-balanced residuals and moist variables [Lines 753–754] “contradicts the hypothesis of stochastic physics exerting a major influence”. They go on to state [Lines 757–759] “The positive trend with time ... again, points to the assimilation uncertainty being dominated by varying observation uncertainties rather than stochastic physics”.	If we understand correctly, this comment regards the complicated interplay between observation uncertainties (perturbations), observational coverage and stochastic physics, especially in the context of moist processes. We agree with the reviewer that this interplay deserves more consideration in the interpretation of the results in Fig. 8. We have added some thoughts about the possible alternative explanation proposed by the reviewer and as well as a connection to the interplay mentioned above raised in other parts of the manuscript.	On the other hand, the generally thin observational coverage in this period probably implies assimilation uncertainty is governed by large-scale processes. The impact of stochastic physics on the spread may therefore not be negligible in an absolute sense, but rather depend on whether observational constraints on the large scales are sufficient for it to become relevant in a relative sense.

Reviewer comment	Response	Changes in manuscript
Have the authors considered the possibility that, in the early record, the variance-balanced residuals will be dominated by larger-scale state-dependent uncertainty growth associated with dry dynamics while, later on, better observations will help constrain these larger scales with the effect that the variance residuals will become relatively more sensitive to smaller-scale rapid uncertainty growth associated with moist physics? Such a possibility does not require stochastic physics to play a negligible role and does not require observation uncertainty (including the perturbation of moisture-sensitive observations) to play a dominant role. This is not to say that the perturbation of observations is the optimal solution. While Isaken et al. (2010) demonstrate (and as alluded to in section 3.1.1), the estimation of background error covariances requires observation perturbations that correctly represent the true observation error covariances, I am convinced that this comes at the expense of increased spread. Machine-learning approaches (for example) to estimating and applying observation error covariances might provide a more optimal solution.		

2 Minor General Comments

Reviewer comment	Response	Changes in manuscript
1. [Line 289] “cf” is used widely to point to confirming information. Isn’t “cf” normally used to indicate opposing information?	We have adjusted our referencing by differentiating between “cf.” when referring to a different (not necessarily opposing) perspective and “see” when referring to the same perspective appearing somewhere else.	

Reviewer comment	Response	Changes in manuscript
2. [Figure 3] Is the unit in the centre panel really $(\log(\sigma_{\text{EDA}}))^2$? Is there really a square here, and writing “EDA” when referring to the residual variance seems odd to me. This is actually a more general issue throughout the manuscript. Please could the authors justify or discuss in the text briefly?	The response-level variance is indeed in units of the response variable squared. The unit of the residuals of a statistical model is the same as the unit of the original response variable. We agree that the variable $\log(\sigma_{\text{EDA}})$ seems odd and have added an explanation at the beginning of the “Results” section.	Our variable of interest throughout will be the logarithmic ensemble spread (standard deviation) of geopotential on 500 hPa (zg500), which we denote by $\log(\sigma_{\text{EDA}})$ to indicate that we take the standard deviation across the EDA members (as opposed to say space, time or WRs).
3. [Line 617] “so interpretation has to be taken with a grain of salt”. There are quite a few colloquialisms used throughout the manuscript. Personally, I find these quaint and amusing, but maybe they will be confusing for the international target audience? For me, at [Line 559], I found it difficult to determine whether “On the other end of the spectrum” is a colloquialism or not.	We have replaced a few instances of such colloquialisms using more professional language.	

3 Minor Specific Comments

Reviewer comment	Response	Changes in manuscript
1. [Lines 108–121] The statements in this paragraph see at odds with each other: environmental or model?	We agree that the paragraph was a little confusing with regard to model and environmental impacts on WCB uncertainty and have clarified.	“Introduction” section

Reviewer comment	Response	Changes in manuscript
2. [Lines 147–149] “instead of seven continuous numerical time series, we simply use one categorical time series of eight categories”. This was a source of great confusion for me. I interpreted this text as defining the covariate x^{WR} . After some time, I now suspect that x^{WR} is deconstructed back to 7 binary variables within the regression(?)	We think this mention confuses more than it helps and have thus decided to cut it, since it is not necessary for the paper. In multinomial regression (where a covariate can take N categorical values), the multinomial variable is replaced by $N - 1$ binary dummy variables with the N th category being absorbed into the offset. We have added a short explanation of this in the methods section.	The WR and the WCB occurrence are both categorical: with L categories, a categorical covariate enters the model as $L - 1$ binary indicator covariates that equal 1 when the respective category occurs at time t and 0 otherwise, with one reference category absorbed into the intercept (treatment coding).
3. [Line 180] “with α increasing by wavenumber.” Is this separate from the spectral smoothing of smaller scales prior to generation of $\mathbf{B}_{\text{EDA}}$ (Bonavita et al., 2012), or should it be “decreasing by wavenumber”?	We have changed the formulation to “increasing with wavenumber” to agree with Hersbach et al. (2020).	
4. [Line 205, Eq. 3] Please give a reference to the second form for the analysis covariance.	We have added a more explicit reference to the relevant Matrix identity.	In the second step of Eq. (3), we have used the Woodbury matrix identity (Higham, 2002)
5. [Lines 218–219] “The contribution of ... in a similar way”. This sentence is unclear.	We have cut the sentence.	
6. [Lines 223–225] I suggest changing the sentence “As uncertain observations ... (Isaksen et al., 2010)” to something like: “As Isaken et al. (2010) demonstrate, the estimation of background error covariances requires observation perturbations that correctly represent the true observation error covariances”.	Agreed.	
7. [Line 299] The word “slightly” does not convey the idea that differences would ultimately saturate at the full amplitude of atmospheric activity, at or before the intrinsic limit of predictability.	We have removed that sentence; it was irrelevant for the study.	

Reviewer comment	Response	Changes in manuscript
8. [Section 4.1] Some context to what the responses and covariates might be would be very helpful.	We provided references to our applications and the variables used therein.	“Methods” section
9. [Line 400] Change “working” to “parametrized”?	Changed “working” to “estimated covariance”	Because both coefficient sets are estimated from the same data, the covariance between them is available analytically (from the design matrices and the estimated covariance, or its HAC counterpart), so the uncertainty of the difference follows by the same linear propagation as for a single EMM
10. [Line 406] Change “levels” to “EMMs”?	We hope changing “levels” to “categories” made this more clear. The categories always differ (in non-trivial cases); it is their effect on the response that is measured by the EMM and, by extension, σ_c .	
11. [Line 412] Change “on the far side of zero” to “outside this interval”?	Agreed.	
12. [Section 5.1] A key question: What happens at the break-points? The β change and/or the parameters defining the residual covariance matrix change? The reader would value this kind of insight.	We believe this is made clear in the “Methods” section (4.2). The index n marks the segment whereby one can see that β_n and σ_n^2 change across a breakpoint and the correlation is “interrupted” by the breakpoints, but the ρ stays the same. We have added a clarification there.	Equation (7) does, however indicate the segments are assumed to be uncorrelated, i.e. the breakpoints break the autocorrelation “chain”.
13. [Line 472] What is the rationale for applying a minimum segment length?	We have added a justification.	We apply such a minimum segment length to prevent severely uneven numbers of data points per segment
14. [Line 474–475] “As mentioned in Sect. 4.3, the confidence intervals for the breakpoints are estimated with a HAC estimator, ...” I do not readily see this mentioned in Section 4.3.	We have added a reference to this in the “Methods” section and elaborate on it in the “Model evaluation and inference” section.	We provide details on the uncertainty quantification for these estimates in App. A

Reviewer comment	Response	Changes in manuscript
15. [Line 495] “Notably, from 1972 to 1978 a new $\mathbf{B}_{\text{cli}}$ was also used” ... but this does not lead to a diagnosed break-point(?)	A breakpoint in 1972 would undercut the applied minimum segment length. That being said, applying a lower minimum segment length does not lead to the algorithm converging on a segmentation with two breakpoints in those years.	
16. [Line 512] What is meant by “The pronounced seasonality in the time series becomes visibly clearer with increasing time”? Increasing time backwards or forwards? I see a strong seasonal cycle in the first segment, then no clear trend(?)	What we tried to say is that the “assimilated” seasonal cycle appears more constant and regular as the time series goes from 1940 to 2024. This refers not to the magnitude of the fitted time series. We have made this more clear.	The pronounced seasonality in the “assimilated” time series becomes visibly more regular as it approaches recent years.
17. [Line 518] “seasonality is a property of the system itself”. What is meant here by “the system”? The data assimilation system, the earth system?	The assimilation system. This has been clarified in the manuscript.	
18. [Line 528] “without assuming the seasonal cycle as a covariate”. On reading for the first time, it was still unclear to me what the remaining covariates were in this case.	We have made more mentions to what the covariates are throughout.	
19. [Lines 539–541] “Crucially, these varying amplitudes are tied to the heteroskedastic nature of the time series; it would therefore be inconsistent to hold the seasonal coefficients constant across breakpoints while allowing residual variance and all other covariates to change”. I don’t quite follow here; surely the heteroskedastic nature is accounted for in the specification of residual variances?	This was a genuine error. As the reviewer correctly pointed out, the heteroskedasticity regards the error structure rather than the covariance structure. We have cut that point and adjusted the sentence.	Although the per-segment seasonal cycles converge towards the whole-period estimate, suggesting one could simplify to a common cycle, the variations in amplitude and phase between segments are significant—more so on daily time scales for individual grid points; it would therefore be inconsistent to hold the seasonal coefficients constant across breakpoints while allowing residual variance and all other covariates to change.

Reviewer comment	Response	Changes in manuscript
20. [Figure 2] The lack of any appreciable $k = 1$ annual cycle in segment 2 is curious. Do the authors have any explanations?	We agree that the minuscule seasonal cycle in segment 2 is curious. We have not investigated this behaviour more thoroughly and cannot think of any explanations apart from the general noisiness of the data.	
21. [Section 5.2] From the beginning, it would be good to know if this section combines the whole period, or represents a single segment, or differences between the beginning and end (as in Fig. D2)?	We have clarified this point.	For a first glimpse into the results for the entire time series as opposed to individual segments, we report the adjusted coefficient of determination R_{adj}^2 across the investigated region together with its decomposition (see App. A).
22. [Line 553] Change “is of due” to “is due”(?)	Agreed.	
23. [Figure D2] Why do the authors not plot $\log(\sigma_{1940}/\sigma_{2024})$, which would give the order of magnitude of the difference?	We show the difference instead of the quotient because the spread is already on a log scale. That way, the difference allows for statements like “The spread has dropped by roughly two orders of magnitude”. We assume the reviewer meant to suggest to show $\log(\sigma_{\text{EDA}}^{1940})/\log(\sigma_{\text{EDA}}^{2024})$, because the quantity they suggested $(\log(\sigma_{\text{EDA}}^{1940})/\sigma_{\text{EDA}}^{2024}) = \log(\sigma_{\text{EDA}}^{1940}) - \log(\sigma_{\text{EDA}}^{2024})$ is exactly what we plotted. If one were to plot that quantity, one could discern by which factor the logarithmic spread fell, which we would argue is less natural than the perspective our plot provides. We concede that this alternative has its merits and provide the quantity as Fig. ??.	
24. [Line 555] Change “Along our” to “Along with our”?	We have replaced that expression.	

Reviewer comment	Response	Changes in manuscript
25. [Lines 562–563] Change “Finally, the scale of the spatial heterogeneity is considerably smaller than the level of heterogeneity in the scale of the total response-level variation” to “Finally, the scale of the spatial heterogeneity OF RESIDUALS is considerably smaller than THAT of the total response-level variation”?	We have clarified that.	Finally, the scale of the spatial heterogeneity in V_{res} is considerably smaller than the level of heterogeneity in the scale of the total response-level variation as can be judged from the color scales in the center and right panel in Fig. 3.
26. [Figure 4 caption] Change “95 % confidence intervals for the mean estimates” to “95 % confidence intervals for the FITTED estimates”?	Agreed.	
27. [Line 597] Change “1.5 orders of magnitude higher for the 1940s” to “1.5 orders of magnitude higher for the Gulf Stream location in the 1940s”?	Agreed.	
28. [Line 602] “... data from the Gulf Stream location remains rather constant”. The curve for this region seems to decline too in Fig. 4. Possibly the authors are talking about data quantity(?)	We are referring to the two recent decades. We have added a clarification.	In this era, with a mature and mostly satellite-based observation system, the differences in the uncertainties at the two grid points become a marginal fraction of their absolute uncertainty. Nevertheless, the two time series do have distinct statistical properties as indicated by the different model estimates for the mean. In particular, the assimilation uncertainty at the Arctic location decreases on the time scale of years to decades whereas the assimilation uncertainty from the Gulf Stream location appears to remain more constant.

Reviewer comment	Response	Changes in manuscript
29. [Lines 604–606] “Consistent with this, we hypothesize that the Arctic grid point is more affected by this (the onset of nearby radar rain-rate measurements around this time appears not to affect the response; cf. Fig. 2 in Soci et al. (2024))”. I am struggling to understand this sentence; maybe there is a double negative involved?	We have removed the tangent to avoid confusion.	
30. [Line 630] “As expected, the PSD peaks at the low-frequency end for both grid points”. Is this “expected” earlier in the manuscript text or appealing to previous studies?	Expected from the shape of the data; e.g. Fig. 1. We have added a reference.	
31. [Lines 631–633] Change “The models also clearly erase local maxima at annual and semi-annual frequencies as well as power at a frequency of four months, though the original data features no prominent peak there” to “The models also clearly erase local maxima at annual and semi-annual frequencies; at four months the original data features no prominent peak there”?	Removed the mention of the four-monthly frequency.	
32. [Line 642–643] “retrospectively justifying the implementation of three harmonics in the modelling of the seasonality”. Does the fact that the residuals at the annual and semi-annual periods drop quite a long way below their spectrally-adjacent values indicate that the annual cycle is still not well catered-for with the three harmonics?	The dips in the residual spectrum at the annual, semi-annual and four-monthly periods are the expected consequence of fitting harmonics at exactly these frequencies. The fit removes not only the seasonal signal but also the small share of stochastic background variance that projects onto these frequencies, and the residual spectrum falls below its spectral neighbourhood. An insufficient number of harmonics would instead show up as residual peaks, e.g. at 3 months for a missing fourth harmonic.	

Reviewer comment	Response	Changes in manuscript
33. [Lines 656–657] “Since we assume the segment-specific heteroskedasticity is an artifact of the observation and assimilation suite”. What is meant by “segment-specific”? The residuals will identify heteroskedasticity within and between segments(?)	We have adjusted the formulation (now in the “Methods” section).	Since we assume the differing error variances per segment are caused by differences in the observation and assimilation suite, we rescale the residuals by the respective estimated error standard deviation in each segment, without changing the overall variance across the whole time series.
34. [Line 660, Equation 20] There is a confusing double use of n here. Maybe use n' in the summation?	Agreed.	
35. [Line 676, Equation 21] Maybe this short-hand equation requires a little more explanation, since x^{WR} is categorical? As mentioned before, everything pointed me in the wrong direction on first reading, because x^{WR} was originally defined a sole covariate with 8 “levels”. Please see my major comment 1.	We hope this concern is addressed by our reworked “Methods” section.	
36. [Lines 680–681] “(i.e. we added x^{WR} to the model in 19)”. Again, please see my major comment 1.	See above.	
37. [Lines 683–684] “Importantly, all three representations yield very similar spatial patterns, ...”. Differences between figures 7, D5 and D6 are almost imperceptible, and I would suggest removing two of these. This is one example of where the paper could be reduced in length.	We have reduced the paragraph to a sentence and pushed the according figures to the supplement.	These spatial patterns are robust to the choice of diagnostic: composite means with block-permutation significance and WR effects estimated within the first-stage model on the original response scale yield very similar results (provided in the Supplement).
38. [Line 701 and 704] Maybe replace “isohypses” with “contours”?	We prefer to keep isohypses, which is more specific.	

Reviewer comment	Response	Changes in manuscript
39. [Lines 703–713] These two paragraphs were difficult for me to understand. One issue was the phrase [Line 709] “which variables attenuate the WR signal”; this meant little to me. Later, I came to understand (I think) that the attenuation discussed relates to the shift in variance from the weather regime covariate to the more physically-meaningful covariates, which thus help explain the WR partition.	We have explained what is meant by attenuation more thoroughly. We have also elaborated a bit on the point discussed in the final paragraph of the section.	In the following sections, we use the WRs as a handle to test these hypotheses: after screening possible explanatory variables in Sect. 5.3.1 & 5.3.2, we quantify which of them reduce the effect of the WRs upon inclusion in a statistical model, i.e. which variables “attenuate” the WR signal (Sect. 5.3.3).
40. [Lines 720–721] “Firstly, the observation uncertainty ingested in the perturbations varies not only on the long time scales isolated by means of our first-stage models”; as a reminder, maybe insert after this, “applied to monthly-mean fields”?	This refers to the first-stage models on daily time scale. We have clarified.	The observation uncertainty ingested in the perturbations varies not only on the long time scales isolated by means of our grid-point-wise first-stage models, but is also state-dependent.
41. [Lines 722–723] “atmospheric moisture and the presence of hydrometeors in particular exert a state-dependent effect on the observation perturbations. Since zg500 is an integrative field and radiances and humidity measurements are important observation types”. Maybe add “humidity-sensitive” before “radiances”?	Agreed.	

Reviewer comment	Response	Changes in manuscript
42. [Lines 724–725] “Perturbations for humidity observations are not applied to SYNOP and TEMP measurements, but are among the measurements with the longest history (Soci et al., 2024)”. Please check this. I think that SYNOP 2m relative humidities were perturbed(?)	We checked the technical documentation for the relevant IFS cycle (https://doi.org/10.21957/9phrx9eu), where in section 2.6.2. “Prescribed observation errors” it says “Relative humidity observation error RH_{err} [...] is prescribed only for TEMP and SYNOP data.”. We have also made the statements more clear.	We argue that this influence has to be varying in time corresponding to a developing observation system. In earlier periods, humidity observations were dominated by TEMP and SYNOP measurements, but those have prescribed rather than derived (state-dependent) uncertainties (Soci et al., 2024). More modern instruments like satellite- or aircraft-based ones do feature derived uncertainties and have become more dominant with time.
43. [Line 732] “(across grid points)” ... and days within the year? Maybe I am missing something here?	The shading in the plot indicates variation across grid points. For each grid point and each year, all daily values of that year at that grid point are used to calculate one estimate for the correlation coefficient.	
44. [Line 732] “implies” doesn’t seem like the right word here?	Agreed.	
45. [Line 752] Change “weak constrain” to “weak constraint”?	Agreed.	
46. [Line 762] How is the layer-average relative humidity calculated?	We have clarified.	Finally, we also include the mean relative humidity across all pressure levels available in ERA5 between 500 and 850 hPa to capture the variable moisture availability in regions most relevant for large scale ascent and convection.

Reviewer comment	Response	Changes in manuscript
47. [Lines 777–828] On first reading, this long discussion on the gradient and Laplacian of the geopotential seemed quite complicated and more of a distraction than informative. Maybe this could be summarized more?	We have shortened it and pushed most of it to the appendix.	Appendix “Interpretation of the gradient and Laplacian covariates.
48. [Line 830] “We also provide maps of correlation coefficients”. This is a strange way to start a subsection.	Agreed and changed.	The correlations in Fig. 8 are medians across grid points and may therefore conceal pronounced regional contrasts, which matter for interpreting the grid-point-wise second-stage models. Figure 9 shows the correlation coefficients at each grid point.
49. [Lines 834–836] “Why exactly a higher cloud cover in upper tropospheric layers is associated with lower assimilation uncertainty is outside the scope of this study – our focus is also more on the mid-latitudes than on the polar regions”. Possibly this polar cloud is shallower, longer-lived, and less associated with less-predictable latent heating?	We are not so sure whether this is a possible explanation. Such a mechanism would perhaps entail a smaller or negligible correlation but not a negative one, right? We think large-scale flow preconditioning might be possible factor. We prefer not to speculate about this aspect as we have not performed a dedicated analysis.	
50. [Line 872, Equation 24] Are the estimated marginal means (EMMs) internally calculated in variance space? (I think that would be better than in standard-deviation or log-standard-deviation space).	We are not completely sure what the reviewer means, but we understand the question as whether the models behind the EMMs should use the variance σ_{EDA}^2 as the response instead of $\log(\sigma_{\text{EDA}})$. We use $\log(\sigma_{\text{EDA}})$ as the response because the spread is positive and strongly skewed. A linear model for σ_{EDA}^2 would violate the model assumptions much more, would be dominated by a few high-spread events and could even produce negative variances.	

Reviewer comment	Response	Changes in manuscript
	The choice should not change our results much. If $\log(\sigma_{\text{EDA}})$ is roughly normal with residual variance s^2 , then $\log \mathbb{E}[\sigma_{\text{EDA}}^2] = 2 \mathbb{E}[\log(\sigma_{\text{EDA}})] + 2s^2$. If s^2 does not depend on the covariates, effects on the mean variance are simply twice our estimated effects, and patterns, significances and attenuation ratios stay the same. We checked that the residual variance differs only slightly between regimes.	
51. [Lines 882–886] “The covariates we add are local, have a finer structure and are more directly linked to physical processes, which is why they don’t ‘compete’ with the WR time series for signal strength, but absorb it to the extent they share variance with it”. What is meant by “absorb”? I would have anticipated that the additional covariates would compete to some extent with the WR. After all, the WRs exist because they are associated with different dynamical/physical processes(?)	We have adjusted the formulation.	The covariates are local, have a finer structure and are more directly linked to physical processes than the WRs, so we expect them to absorb a major part of the WR effect.
52. [Line 912] Change “deduct” to “deduce”?	Agreed.	
53. [Line 929] What is the analogous equation to (25) for the “leave-one-out” strategy?	We have added the equation.	(17)
54. [Lines 932–933] “Results from ‘add-one’ models mentioned above would look virtually identical to fig. 9”. Fig. 9 shows correlations, not $\Delta\sigma$ (?)	The sentence was confusing and a little speculative so we removed it.	

Reviewer comment	Response	Changes in manuscript
55. [Lines 937–938] “A large part of the variability in the uncertainty that is projected onto the WRs clustering is due to the uncertainties’ correlation with the geopotential gradient itself”. I think that the approach being used here aims to estimate the amount of the WR variance that can be accounted for by the various other covariates(?) Could some text like this be used to motivate the methodology?	We have added some motivation in the “Methods” section.	Section “Attenuation analysis”
56. [Lines 941–944] “The large and significant patch of negative values across northeastern North America and along its coast imply a similar conclusion. This is because the WR patterns exhibit minuscule variability in the zg500 gradient in that region whereas the region is generally quite variable in that respect”. I don’t fully understand the discussion around the negative attenuations. Is another approach to mask out regions where the WR patterns have small signal? This would certainly remove northeastern North America.	We have elaborated on the explanation of the negative attenuation.	Positive values indicate that the added covariates absorb part of the WR effect. Negative values are possible as well: covariates that vary strongly within regimes but little between them can mask between-regime differences in $\hat{\varepsilon}^*$ as long as they are not accounted for.

Reviewer comment	Response	Changes in manuscript
57. [Lines 951–952] “We consider it also a further argument against the Laplacian’s effect being mediated through its role in QG dynamics, since QG factors certainly have varying effect magnitudes across the domain”. To the extent that the Laplacian of the geopotential resembles minus the geopotential itself, would this not explain why the Laplacian is able to account for a lot of the WR variance?	We thank the reviewer for this interesting thought. The Laplacian equals the negative geopotential (times a factor) only for a single spatial scale (for a spherical); for a field comprising many scales, it strongly emphasizes the smaller, synoptic scales, whereas the geopotential itself is dominated by the planetary scales. Nevertheless, since the Laplacian is a linear operator, the regime-mean Laplacian equals the Laplacian of the regime-mean geopotential, which for the smooth, large-scale regime composites does resemble the negative regime anomaly. Part of the attenuation by the Laplacian may therefore reflect its similarity to the regime patterns themselves. However, if this were the dominant mechanism, we would expect the attenuation to be largest where the regime patterns differ most, as is the case for the gradient (Fig. E5; formerly D10). Instead, the attenuation by the Laplacian is fairly uniform across the domain, which we interpret as evidence for a more general effect. We have added a few sentences on this to Sect. 5.3.3.	One might suspect that the Laplacian simply mimics the regime patterns: averaged over a regime, the Laplacian equals the Laplacian of the regime-mean geopotential, which for smooth, large-scale patterns resembles the negative of the pattern itself. If this were the main mechanism, however, the attenuation should be largest where the regime patterns differ most, as is the case for the gradient. Instead, the Laplacian accounts for a considerable part of the signal in a regionally constant manner, which speaks in favor of a general effect. We consider this also a further argument against the Laplacian’s effect being mediated through its role in QG dynamics, since QG factors have varying effect magnitudes across the domain.
58. [Lines 963–964] “... much of that uncertainty commonly attributed to model formulation rather than to the initial state ...”. It is something inherent in the real world rather than just sensitivity to model formulation(?)	We have added a mention.	As outlined in the introduction, WCBs are well established as amplifiers of forecast uncertainty, which is commonly attributed to model formulation and intrinsic instabilities rather than to the initial state.
59. [Lines 978–969] “our target variable”. This is a bit vague; can you be more specific?	We have adjusted the term to “response variable”	

Reviewer comment	Response	Changes in manuscript
60. [Line 977, Equation 26]. I appreciate that inflow, ascent and outflow are discussed above, but it could be worth defining what is meant by superscripts “inf”, “asc”, and “out”.	Agreed.	
61. [Lines 994–995] “Drawing attention to the differences in the range of values covered by the colour bar of fig. 12 compared to the colour bar in fig. 7”. Maybe some of the signals seen on the eastern flanks of WR troughs relate to the WCB signals seen here?	We have added a mention	We find it quite plausible that some of the signal seen in Fig. 7 is caused by WCBs—in particular at the eastern flank of troughs.
62. [Line 997] “lower frequencies and spatial locality”. Spatial locality: yes. I don’t follow the lower frequency argument though(?)	We have clarified the formulation.	Some of the differences in magnitude arise because WCB occurrence is a strict and local criterion: it selects those time steps at which a dynamically active air stream is present at the grid point itself, whereas a WR label describes the large-scale flow and also covers many locally quiet situations, which dilutes the contrast between categories.
63. [Line 1005] Change “background forecast ensemble forecasts” to “background ensemble forecasts”?	Agreed.	
64. [Figure 13] Please include the other (overlap?) colours in the key.	We have adapted the figure.	
65. [Lines 1017–1018] “The procedure yields nine binary masks per WR”. Is this $3 \times 3 = (\text{inf}, \text{asc}, \text{out}) \times (0\text{ h}, 12\text{ h}, 24\text{ h})$?	Yes.	
66. [Line 1025] Change “source” to “inflow” for consistency?	Agreed.	

Reviewer comment	Response	Changes in manuscript
67. [Lines 1031–1032] “sufficient for meaningful estimation of the WR specific influence of WCBs on assimilation uncertainty”. This last section, bringing WCBs and WRs together, adds further complexity and more uncertainty to the analysis. I wonder if it adds enough to the study to warrant its inclusion(?) By the way, Figure 14 does not seem to be mentioned in the text.	We have shortened the manuscript in other places, we consider less interesting and hope this justifies keeping this final analysis. We have also added a few sentence to highlight aspects we find relevant additions to the paper’s main message. We have also added reference to Fig. 14. We are happy to cut the passage should the reviewer insist.	Section “Warm conveyor belts”
68. [Lines 1104–1106] “Rather, the presence of moisture exerts a state-dependent (positive) effect on the assimilation uncertainty through elevated observation uncertainties due to both increased observation perturbations and decreased observational coverage – in particular in regions with prevalent convective activity”. Please could the authors re-visit this conclusion in the light of their reply to my major general comment 4? Maybe this conclusion needs to be a little more nuanced?	We have adjusted the relevant paragraph and want to thank the reviewer again for this valuable insight. It has both strengthened our manuscript and elucidated a question we have long puzzled over.	On the other hand, stronger observational constraints especially on the large scales shift the assimilation uncertainty to smaller scales where moist processes are more relevant. This would also explain why stochastic physics associated with moist processes do not lead to assimilation uncertainty in the observation sparse early periods even though they are active throughout.

Reviewer comment	Response	Changes in manuscript
69. [Lines 1112–1114] “On a more general level, the results presented herein point to short time scale assimilation uncertainties being driven by observation specific aspects more than physics of short-forecast error growth”. Again, this conclusion might need to be more nuanced following the authors’ reply to my major general comment 4? The authors do go on to state that “While regions and times associated with elevated uncertainties may relate to both – namely in convectively active regions – it has been shown, that this is not always the case” but this confuses me, since the convectively active regions are precisely what we are interested in(?) Could the authors state explicitly the regions and times where it is not the case?	We have adjusted the paragraph.	On a more general level, the results presented herein point to short-time-scale assimilation uncertainties being driven in large part by observation-specific aspects in addition to short-forecast error growth. While regions and times associated with elevated uncertainties may relate to both—namely in convectively active regions like WCBs—the differences in predictability associated with WRs do not translate to higher or lower total assimilation uncertainty in WRs.

Reviewer comment	Response	Changes in manuscript
70. [Lines 1115–1116] “We therefore argue considering flow-dependent assimilation uncertainty is vital, yet frequently overlooked in studies concerned with error growth or ensemble forecast uncertainty, especially from a WR perspective”. I totally agree and wonder if this is an avenue that the study could draw more conclusions from.	We are glad the reviewer shares this view and have added a sentence elaborating that point.	How exactly the regionally specific uncertainties are amplified in a way that lead to these predictability differences is not yet fully understood and we argue considering flow-dependent assimilation uncertainty is vital, yet frequently overlooked in studies concerned with error growth or ensemble forecast uncertainty, especially from a WR perspective. By quantifying where and how strongly assimilation uncertainty differs, this study provides the reference needed to prevent confounding insights about error-growth processes with mechanisms actually attributable to the assimilation system.
71. [Line 1152] Change “ $e_B = T$ ” to “ $e_B = \tau_B$ ” (?)	The block with indicator “B” is the last block, so the index signifying its end e_B is equal to the index signifying the end of the total time series T .	
72. [Line 1155–1156] “subject to the constraint that adjacent blocks in each permutation do not contain the same category”. I think I am not understanding something here; since these are permutations, how could adjacent blocks contain the same category?	In principle, a permutation can also lead to blocks of the same category being adjacent. This is because the permutation works (roughly) by randomly drawing from the set of blocks one after another without replacement. Imagine you stack multiple decks of cards that are initially ordered. If you shuffle them and then redistribute, you may end up with multiple queens of hearts directly one after another.	
73. [Line 1158] Change “A2” to “A1”?	Agreed.	

Reviewer comment	Response	Changes in manuscript
74. [Figure D4] Are plots centred around the annual-mean values?	The plots show only the effect of the seasonal covariates, which by construction have zero mean; so yes.	

4 Figures

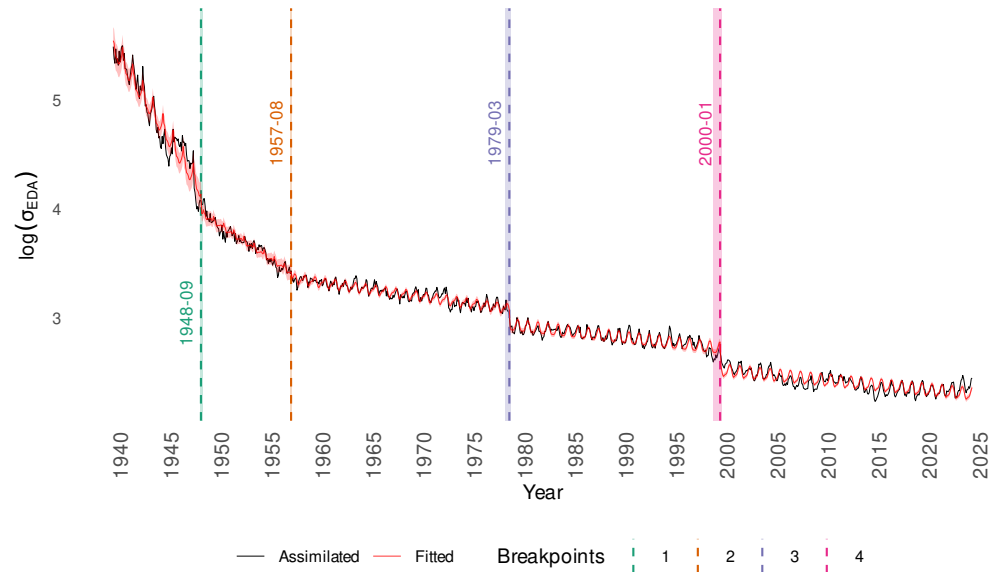


Figure 1: Same as Fig. 1 in the manuscript, but based on a segmentation with four breakpoints instead of five.

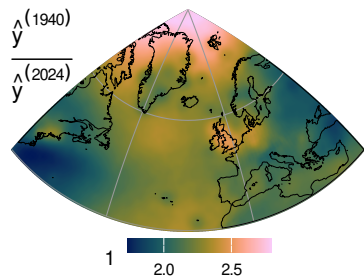


Figure 2: Same as Fig. E2, but showing the relative instead of the absolute difference.