Answers to the comments made by reviewer Peter Pfleiderer to the article "*A flexible methodology to evaluate natural variability in ClimaMeter*"

May 2026

First and foremost, we would like to thank the two reviewers for taking the time to carefully read the manuscript. Their comments are the result of careful consideration and are very insightful. We have done our best to address all of their remarks, and we believe that our work has truly benefited from this review and that the manuscript has improved significantly.

In this document, the reviewer's comments are shown in blue and our responses in black. All line numbers refer to the submitted version of the article.

The authors present an update (ClimaMeter2.0) of an analogue-based extreme event attribution (EEA) framework ClimaMeter that relies on observations (reanalysis). It is an important method complementing other EEA methods that often rely on climate model simulations. The updates intend to fix flaws in the previous version of ClimaMeter by explicitly choosing a hypothesis based on a trend analysis and by weighting modes of natural variability based on the importance for the studied event. The methods are well explained and three examples help to understand the methodological changes. The paper is well written and structured. The synthesis of different parts of the analysis makes the interpretation of the results challenging. Before publication, some more reflection and evaluation on the importance of different parts of the analysis for the final attribution statement and the synthesis of different would be helpful.

I find that the tests (quantile regression, similarity in modes of internal climate variability, and difference signal in the analogues) are interesting for the interpretation of an observed extreme event. I also understand that the authors want to present a highly synthesized result for communication purposes. But I think that the synthesis of the three tests is challenging and needs some more testing and reasoning.

Major comments:

1. For instance, I find the interpretation of the results of the quantile regression problematic. The authors use a the quantile regression to set the confidence level for a test in similarity between modes of internal variability in factual (F) and counterfactual (CF) climate. Over shortish periods, ENSO, PDO and AMO can influence the results of the quantile regression. For instance, the AMO index increases over the reanalysis period. There could be cases, where you detect a significant trend in your extreme index due to this "trend" in AMO (if the event is influenced by AMO). Therefore I'm wondering, what level of evidence the statistical trend analysis brings and how much one should rely on this trend analysis to define the confidence level for later tests in the methodology. Could there be cases of circular reasoning? AMO influences event and AMO is increasing in the studied period $\Rightarrow$ strong statistical trend due to AMO $\Rightarrow$ less strict test when comparing AMO values between analogues $\Rightarrow$ ACC intensified event

   In ClimaMeter 2.0, we assume that a significant trend in the quantile regression is ascribable to climate change. We know that this is a strong and simplified assumption, and we emphasize this in the limitations of ClimaMeter 2.0. At the same time, we believe that the analysis of unconditional trends in hazards is a reasonable measure of the local effect of anthropogenic climate change. For example, in the case of temperature, this is done regularly. The increase in AMO during the instrumental period does not preclude the conclusion that the temperature trend is due to human influence. However, the point raised here is very interesting, and it would be worthwhile to separate the relative contributions of the three modes of variability and the ACC in determining the trend in hazard.

   We will stress this concept in the manuscript, adding at line 601: "*Nevertheless, we believe that unconditional trends in hazards provide a reasonable starting point for measuring the local effect of ACC. As discussed above, this methodology is conceived to be flexible and easily adaptable to new choices of the null hypothesis that may be proposed in future work. For example, a potential improved approach could examine whether the modes of variability under consideration play a role in the long-term trends, and, if so, separate their contribution from that of ACC.*"

2. More generally, I would find it interesting to check how relevant the individual tests are for the final attribution statement. If it would turn out, that the final attribution statement (gauge value) heavily relies on the quantile regression I would suspect the results to be very similar to probabilistic EEA applied to observations only (GEV fit with the location varying with GMST). In my view that would be worth commenting as the analogues methods is seen as an important complementary approach to the probabilistic EEA. "Hiding" the result from the analogue comparison because of the trend analysis would be a shame in my opinion.

   In ClimaMeter 2.0, we introduce two analyses that do not depend on analogues: weight calculation and trend analysis in specific quantiles of the local hazard. Based on the tests we have conducted, it is indeed the test of the null hypothesis (performed via the quantile regression, as detailed in section 4.2) that plays the most important role in determining the value of the gauge 2.0. In fact, among the various formulations for the new natural variability gauge, we tested one that includes weight calculation but does not test the null hypothesis and therefore retains the same indicator function as ClimaMeter (as expressed in Eq. 7). This formulation did not yield satisfactory results in pre-industrial experiment, as it did not substantially reduce the fraction of events identified as influenced by climate change. We will better stress this in the article, by adding at line 584: "*In fact, we tested a version of the gauge that includes only weight calculations but not trend analysis, and this version did not yield satisfactory results in the pre-industrial experiment, as it did not significantly reduce the percentage of events identified as being influenced by climate change (not shown).*"

As suggested in this comment, such step is, in some ways, comparable to fitting a generalized extreme value distribution, the first step in probabilistic attribution approaches: while we do not assume that the variable follows a theoretical distribution, we nevertheless make statistical assessments of its extremes, independent of circulation analogues. At the same time, the framework remains firmly grounded in the analogue methodology and continues to complement probabilistic methods. Specifically, a fundamental difference between the two lies in the modeling of natural variability. In ClimaMeter 2.0, as in ClimaMeter, it is decomposed into recurring sea surface temperature patterns, and the relationship between these patterns and the occurrence of analogues is investigated. In this sense, attention is maintained on the physical process leading to the events, a characteristic of the storyline approach within which the analogues methodology is embedded. In the World Weather Attribution protocol (the most well-known probabilistic framework for rapid attribution), however, natural variability is primarily considered as a source of sampling uncertainty that needs to be quantified. Both approaches have advantages and disadvantages, and provide valuable and complementary information. The point raised here is very interesting, and we will refer to it in the manuscript by adding at line 580: "*It is worth noting that the characterization of unconditional trends in hazard can be seen as a feature shared with probabilistic attribution methods. These methods involve, among other steps, fitting a statistical distribution that varies with a covariate describing global warming to a long time series of observations (Philip et al., 2020). With the quantile regression, we indeed perform a statistical assessment of extremes: we evaluate whether events located in the upper tail of the distribution have become more or less intense over time. However, this unconditional analysis serves as the basis for the conditional analysis of changes in analogues. In fact, it leads to the choice of the significance level to be used when comparing sets of counterfactual and factual natural* *variability indices for analogue dates. So, even though the features we are introducing are non-conditional, ClimaMeter 2.0 remains firmly grounded in the analogues methodology and continues to complement probabilistic methods.*"

3. From the tests in long piControl runs, I do get the impression that the quantile regression is quite relevant in ClimaMeter2.0. I think that the piControl test was important to identify a flaw in ClimaMeter and it nicely shows that problems can be solved with ClimaMeter2.0. However, the approach only tests how often you get false positives. I think that an evaluation with in single model large ensembles would be very informative. Applying the method to single simulations of an ensemble and to evaluate well understood kinds of events (using the information of the ensemble mean) would be very useful.
   The use of single-model large ensembles would indeed be of great interest, as it would give us the opportunity to study more extreme events and, more generally, to have a more complete sampling of natural variability. At the same time, in the pre-industrial experiment we propose, 2,000 years of simulation are available. This long time span provides a statistically significant sample. Furthermore, our aim was to verify that ClimaMeter 2.0 is flexible and adaptable to different climate states. In this regard, we felt it was more relevant to use several climate models – which are associated with different representations of natural variability – and to verify that ClimaMeter 2.0 performs well across these models.

4. I'm wondering whether only showing three simple statement, one for each of the tests, would be more interesting for the lay persons. For example for the European heatwave 2023 you could say something like: "Heat extremes increase in this region" (quantile regression ⇒ fig. 7d) "Similar events were 0.86C cooler in the 1960s" (Analogues ⇒ fig. 7c) "Slow fluctuations in the climate system over the period 1950-now have influenced heat waves in this region" (fig. 7g) The last statement already gets a bit technical and I understand why you would want to avoid it. But I think that a few iterations with people outside of the field would lead to a wording that is understandable, interesting and correct. If you still want to show a synthesized main result I would strongly suggest to add such statements additionally.
   The communication aspect is indeed very important. Based on the important suggestions made in this comment, we have added the following statements:
   line 484: "*By limiting the interpretation to the target region, Fig.7c indicates that events similar to the heatwave were on average 0.86°C cooler during the counterfactual period.*"
   line 498: "*As shown in Fig.7e, the ClimaMeter gauge gives a values of 35%. The interpretation of this result is that changes associated with heatwaves with similar circulation patterns on the spatio-temporal domain of the 2023 event (omega blocking on the domain [18°W; 20°E] x [55°N;*

33° N] for a duration of 8 days) are likely driven also from natural variability."

line 504: "*This results suggests that in central and northern France, heat waves of similar or greater magnitude than the one of September 2023 are becoming more intense in recent climate because of climate change.*"

We would, however, like to emphasize that this is a methodological study not intended for the general public. The analyses published within the ClimaMeter, on the other hand, follow a standard protocol that makes them easier for the general public to interpret. For each event, for example, a press summary is provided—that is, three sentences that convey three key aspects of the analysis. Among these are statements regarding changes in analogues between the two periods and the role of natural variability, which reflect the points made in this comment.

5. Although I find the reasoning of the methodology quite convincing in the sense, that I don't know how a synthesis targeting one single result value could be done better, I'm a bit skeptical whether such as synthesis is a reasonable representation of the results.

   We understand the concerns raised in this comment. Indeed, it is very challenging to represent the complex interaction between natural variability and climate change in the occurrence of extreme events using a single value. At the same time, the need to provide a result that is easy for the general public to understand is dictated by ClimaMeter's communication objectives. The methodology we propose represents an improvement and a more refined representation of natural variability. Furthermore, it is designed to be easily adapted and updated to any improved version.

6. The examples are very useful to understand the methodological changes. The interpretation of the final result is however missing. A brief discussion on how the results can be interpreted in terms of physical mechanisms would greatly improve the confidence in ClimaMeter2.0. In the current state, the examples leave me with questions:

   Heatwave 2023: Looking at fig. 7c I would have said, that only very few grid-points within the region of interest show a significant difference in temperatures. Is there a threshold for the significance of delta T? We know quite well, that heat waves in this region have intensified substantially and actually more than in other regions. This is due to circulation changes over the period of interest (forced or not), aerosol reductions and soil-moisture changes. How do you interpret, the result of ClimaMeter 2.0 saying that the heat wave in 2023 was not influenced by ACC? Depending on how ACC is defined, the reduction in aerosols over the region could be seen as a reason for an intensified heatwave in this region where ACC did not play an important role. Since you do not treat aerosols explicitly, I'm however wondering how the method comes to this conclusion and how this should be interpreted. Does it mean that for heatwaves coming from this kind of circulation pattern the effect of ACC is negligible?

   Regarding the interpretation of the gauge value, we will add at line 498:

"*As shown in Fig.7e, the ClimaMeter gauge gives a values of 35%. The interpretation of this result is that changes associated with heatwaves with similar circulation patterns on the spatio-temporal domain of the 2023 event (omega blocking on the domain [18° W; 20° E] x [55° N; 33° N] for a duration of 8 days) are likely driven also from natural variability. In other words, since two of the three modes (AMO and PDO) are in different phases during the factual and counterfactual periods, their influence cannot be ruled out.*"

In this regard, the value of Gauge 2.0 is consistent with that of ClimaMeter, and their interpretation is the same.

Concerning the conditional delta in the target region, grid points with significant differences are identified using a bootstrap test. Significant grid points are defined as those that fall more than two standard deviations above or below the mean of the bootstrap sample. The target region was selected to include both the major anomalies associated with the event and the most significant adjacent impacts (the Paris metropolitan area). Although only a minority of grid points exhibit significant conditional differences, this is nonetheless a significant signal. We could have chosen other cities heavily affected by the event, such as Rome and Lyon and the results (specifically the significant changes) would have been different.

As for the possible influence of local forcings on temperature, a key working hypothesis is that we assume significant long-term trends are due to climate change. This is a strong but necessary assumption, which is also adopted, for example, in probabilistic methods (Philip et al., 2020). It follows that, given the current state of methodology, we are unable to attribute changes in hazards to, for example, aerosols, land use change, or soil moisture. The point raised here is very interesting, and it would be worthwhile to verify whether these factors are indeed responsible for the intensification of the heatwave in question, rather than ACC. We believe that this potential investigation, despite its great value, goes beyond the scope of this work. At the same time, it is worth further emphasizing the working assumption just discussed. We will add at line 318: "*It is important to emphasize that this second feature is based on the assumption that a significant trend in the hazard variable is due to climate change. This means that we are unable to assess the effect of local forcings (such as aerosols, soil moisture, and surface roughness) on hazards. This hypothesis is also shared by probabilistic attribution methods (Philip et al., 2020) and must be kept in mind when interpreting the results.*"

Storm Hans: The analogue comparison shows a northward shift of the precipitation with insignificant changes in the region of interest (except the parts at the border of the region). The region where extreme precipitation was observed are not affected by these differences. Why would you say that heavy precipitation was influenced by ACC? It seems that the result strongly depends on the extend of the region of interest.

It is indeed true that there is a dependence on the region chosen as the target, especially for events characterized by strong winds or heavy precipitation, which are less spatially homogeneous than temperature extremes. The target region is selected using a semi-objective approach and is assumed to have a rectangular shape for convenience. This can introduce a certain degree of subjectivity, and it is for this reason that among the future directions of this work we have included the determination of the target region as the grid points where the hazard associated with the event exceeds a certain threshold. For example, we performed this analysis for Storm Hans, as shown in Fig. 1 below. For each grid point, we calculated the 98th percentile of the precipitation distribution over the combined factual and counterfactual period. We then identified the grid points where the precipitation associated with Storm Hans exceeds this percentile, shown by the stippling in Fig. 1b. We selected an area of interest (rectangle in Fig. 1b), which includes the heaviest precipitation and the most significant impacts on land. We defined the target region as the set of grid points within the area of interest with precipitation above the 98th percentile. We repeated the analysis using this target region. This allows us to better appreciate that, in fact, in some grid cells that have experienced extreme precipitation, there is a significant increase in precipitation. The results of the weight calculation and quantile regression change slightly with respect to the analysis presented in the pre-print, precisely because of the different target region. This procedure allows us to select a target region that is less arbitrary and more meaningful. To illustrate ClimaMeter 2.0, we think it is sufficient to select a rectangular target domain, as this simplifies communication. In the future implementation of ClimaMeter 2.0, we believe that selecting the target region using thresholds is critical.

We have further emphasized the need for this improvement, modifying line 627: "*An interesting development could be to automatically select the target region by defining grid-points where the hazard is overcoming a certain threshold. The choice of the target region is critical and the results dependent on it, as it is subsequently used to estimate trends in the hazard variable. Automating this step would represent a great improvement, since it would reduce subjectivity in the choice of the domain. Specifically, this advancement is of crucial importance for events involving extreme winds or precipitation, which can often be very intense on a local scale.*"

Storm Ciaran: Figure 9 suggests that for similar large scale circulation patterns, local wind speeds are higher in F as compared to CF. In terms of conditional attribution one would say that ACC played a role for this kind of circulation type. There is no significant trend which could be to a decrease in frequency of weather patterns producing strong winds or winds being weaker for other types of weather patterns. I thought that with ClimaMeter you were targeting a conditional attribution for the relevant large scale circulation patterns. The synthesized result saying that ACC did not play a role is a surprise to me and to me the gauge does not represent the results shown in other panels of figure 9. I understand how the value of the gauge was obtained. I'm rather wondering whether this

is what ClimaMeter should show us for cases where the trend is insignificant. To me it would make more sense to point towards the IPCC with a statement like: "ClimaMeter finds that wind extremes similar to Ciaran were intensified by ACC even though there is no overall intensification in wind extremes in this region"
That is correct; we agree that the absence of a significant trend in the quantile regression should not invalidate the other results. In the case of Storm Ciaran, we will stress this aspect adding at line 550: "*The interpretation of such result is that in recent times, extratropical storms similar to Ciaran have been characterized by stronger winds (Fig.9c). However, it is challenging to attribute this increase to climate change, as there is no evidence of a sustained intensification of extreme wind events in this region. For this specific event, natural variability seems to be the determining factor leading to the increase in wind speed.*"
When the natural variability gauge points toward 0% (often associated with the absence of a trend in quantile regression), this must be handled with great caution, especially when communicating the results.

7. L35: New simulations are not required for each EEA study. Existing large ensembles or other simulations can be used.
   This has been changed to: "*Although EEA has substantially advanced our understanding of the role of climate change in extreme weather events, conventional attribution studies often require weeks or even months to be completed.*"

8. L39: Reference for WWA is missing: added

9. L71-75: I find this part a bit confusing and I think some clarifications would help. What about forced dynamic changes? I think that for this part of the argumentation, a separation between forced dynamical changes and forced thermodynamical changes would be helpful. Is the overall assumption in this article that dynamical changes are not forced? If that is the case please state it explicitly.
   We will add at line 75: "*In general, the response of atmospheric circulation to climate change remains poorly understood (Shepherd, 2014; van Garderen et al., 2021), observational records do not show clear trends (Deser et al., 2016), and, furthermore, there is not even a sufficient degree of agreement among models (Pfahl et al., 2017). Therefore, since the atmospheric thermodynamics interact with dynamic aspects, the choice to systematically assume that ACC plays a critical role in extremes may be too general.*"

10. L173: You compare the modes of variability of the days from the selected analogues. I would reformulate this and not say "between the two periods".
    This has been changed to: "*When the p-value obtained with the test is smaller than 0.05, the null hypothesis is rejected and the distributions are*

identified as significantly different. The physical interpretation is that, if the mode of variability has a different overall phase between analogues in the two periods, than its influence cannot be ruled out."

I'm also wondering about the interpretation: there could be different cases leading to different modes in the selected analogue days: (i) the whole period has on average different values in these modes of variability (this you can actually check). (ii) analogues of the event occur in different phases of the modes in the factual as compared to the counterfactual climate. (iii) analogues in one period are by chance in different phases of the modes. Would it make sense to discuss these cases explicitly?

The three possibilities outlined in this comment could indeed occur. For example, the phases of interdecadal variability modes such as the AMO and PDO are not fully sampled over periods of approximately 40 years. A difference between the phases associated with analogues in the two periods could, in principle, arise solely from a shift in the mean phase of the mode over the entire period. For this reason, we assign weights to the three modes based on the strength of the teleconnection between the remote mode and the local hazard. Investigating the relationship between modes of variability and the occurrence of extreme events and their analogues is of considerable interest, and the analogue methodology readily allows such analyses. However, this topic is extensive and lies beyond the scope of the present study.

11. L179: I'm a bit confused about the "Influenced by Natural Variability": as I understand the test, you check whether the analogues in factual climate occurred in different phases of the modes of variability as compared to pre-industrial climate. With this you can make a statement about the 30 analogues in factual climate. Can you also make a statement about the observed event ClimaMeter is analyzing? Or how exactly has this "Influenced by Natural Variability" to be interpreted?

    This comment provides an opportunity to clarify an important point. The natural variability gauge refers to changes over time in events similar to the one under consideration. By "similar," we mean "characterized by atmospheric circulation conditions close to the one of the event". We can state, for example: "storms similar to Storm Hans became up to 5 mm/day wetter, and natural variability likely played a minor role in determining this increase." The magnitude of the event and the conditional changes are two distinct entity. What is ascribed to climate change is not the event itself, but rather the changes associated with similar events, namely the analogues. The relationship between the event and the conditional delta determines whether similar events have been strengthened or weakened by climate change, as discussed in section 3.1.

    To better highlight the difference between the event and changes in analogues, we will add at line 192: "*It is important to note that the magnitude of the event and the changes in analogues (Fig.1 step 4) are two distinct entity. The natural variability gauge, given its design, allows us to make* *statements about significant evolutions of the analogues. In fact, it assess the potential roles of natural variability and climate change in determining changes conditioned to the analogues, and not in driving the occurrence of the event itself.*"

12. L219-221: Please also name one of the papers that discusses observed and simulated changes in ENSO.
We will add at line 219 (section 3.2.2): "*For ENSO, there is evidence showing that its amplitude from 1950 to the present has been higher than during the pre-industrial period (from 1400). This suggests that such an increase in variance may have been caused by enhanced greenhouse gas levels (Hope et al., 2017). In addition, there is evidence that human emissions and changes in aerosols have an impact on the persistence of El Niño events, although these effects have compensated each other over the course of the 20th century (Stevenson et al., 2019). However, no clear evidence indicates a recent and sustained shift beyond the range of decadal to millennial variability (Gulev et al., 2021), and there is high confidence that human influence has not yet modulated its variability beyond its range of internal variability (Maher et al., 2018).*"

13. L298: I don't understand the reasoning behind option (b) in the ACC test. Since you already select analogues and control for modes of variability, what would be the fluctuation explaining the conditional difference? To me, this additional test is rather confusing because I thought the aim of ClimaMeter is specifically to do a conditional attribution. I understand the motivation to work on the null-hypothesis that ACC explains everything remaining after correcting for the modes. Maybe you could add an example of a case where this test helps to identify a mismatch between trend and conditional difference and how this can be explained "natural fluctuations of the climate system".
In fact, the idea behind quantile regression is not to identify climate fluctuations or a specific mode of variability responsible for the conditional delta. The goal is to add an unconditional test in order to select the most appropriate null hypothesis regarding the conditional delta. With quantile regression, the question we are seeking to answer is: do analogous events show changes consistent with the local effect of climate change, or do they show a different change that we expect to stem from natural variability? We answer this question by comparing quantile regression with the conditional delta (section 4.2). Consequently, the significance level for the statistical test used to compare the index values associated with the analogues in the two periods depends on this. Furthermore, quantile regression allows us to perform a regression specifically tailored to the intensity of the event under consideration.

14. L394: Beyond improvements to previous versions of the model, is the performance satisfactory? To make this point clearer, we will add at line 393: "*Regarding the IPSL model, Boucher et al. (2020) evaluate the characteri-*

zation of ENSO and state that its seasonality has improved radically since the previous version of the model, while other aspects are quite similar (such as the too small difference in amplitude between spring and winter ENSO events). Bonnet et al. (2021) assess AMO and PDO and find a good representation of their spatial pattern and temporal evolution, consistent with observations. Overall, the authors conclude representation is satisfactory and is the best among the various versions of the IPSL model."

15. L551: "It is important to stress that such statement is only valid for storm Ciaran, its dynamics and associated wind speed, and cannot be generalised to any other storm or event." Is this the case because the trend significance depends on the region? Otherwise I find this statement misleading as this trend analysis does not depend on the circulation types.
The calculation of the gauge value comprises a component that is independent of circulation (trend analysis and weight computation), but also a component that is dependent on analogues (the p-values in Equation (5), derived from statistical tests of the indices on analogue dates between periods). For this reason, the natural variability gauge remains an indicator closely linked to the event under consideration and its dynamic characteristics.

Minor comments:

1. L37: one point to much: corrected
2. L44: "analyzed events" ? corrected
3. L447: which pre-industrial heatwave? The characteristics of the pre-industrial heatwave have been selected to reflect those of the ClimaMeter study of the September 2023 European heatwave, as described in section 5.3.1.
4. L484: degree C (or K): corrected

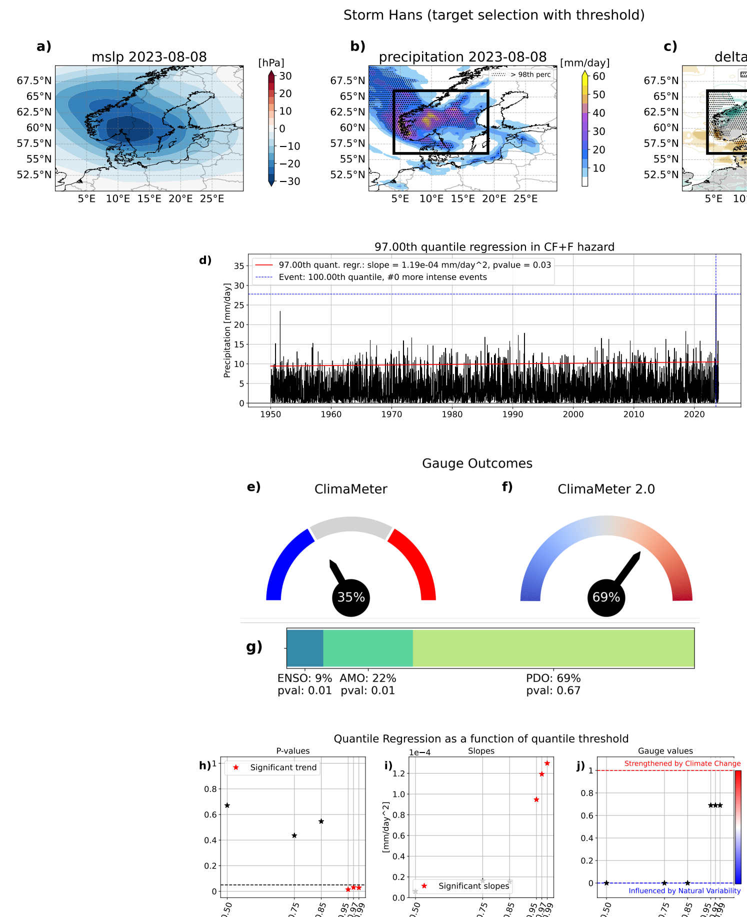

Figure 1: Analysis of Storm Hans. (a) Mean sea level pressure anomalies and (b) precipitation maps associated with the event. (c) Significant changes in precipitation between factual and counterfactual analogues composites. Fields in a-c) cover the analogues domain ([0°E; 30°E] x [70°N; 50°N]). The black square ([4°E; 19°E] x [66°N; 56°N]) in b) indicates the area where most extreme hazard and impacts occurred. Stippling indicates grid points at which the precipitation associated with Storm Hans exceeded the 98th percentile of the local precipitation distribution. The target region is defined as the set of significant grid points within the area of interest represented by the rectangle. d) Time evolution of precipitation spatially averaged in the target region over the union of counterfactual and factual periods. The blue dashed line represents the precipitation associated with the event. The legend indicates the quantile corresponding to the event intensity within the distribution of the target precipitation, as well as the number of previous events with higher intensity. The red line represents the 97^{th} quantile regression of precipitation, and slopes and p-values are reported in the legend. Note that a p-value smaller than 5% is associated to a significant trend, since the null hypothesis of the test is that the regression slope is zero. e-f) Comparison of natural variability gauge outcomes following ClimaMeter and ClimaMeter 2.0 methodologies. g) Weights expressed as percentages for the modes of variability, and p-values obtained from the CvM test at 0.05 significance level between counterfactual and factual set of indices. h-j) Quantile regression for different values of the quantile threshold: 0.5, 0.75, 0.85, 0.95, 0.97, 0.99. i) p-values and j) slopes associated with the regression; k) natural variability gauge value based on the quantile regression methodology, all as a function of the quantile threshold.