Reply to the reviewer comments

Jonatan F. Siegmund *et al.*

We would like to thank both referees for their comprehensive reviews and helpful comments on our discussion paper. In order to address all comments and raised concerns and avoid duplications, in the following we have rearranged both reviews slightly and reduced the full reviews to the most relevant aspects.

COMMENTS OF REFEREE 1

1. *...it would be interesting to see how the magnitude of a climate extreme is related to the shift in flowering dates (in days). This would strengthen the message of the paper ...*

   We thank the reviewer for making this important point. We fully agree that an analysis like the one proposed here would be a valuable additional aspect of the present manuscript. We have followed the corresponding recommendation and have made corresponding further calculations. In our revised manuscript, we plan to accordingly provide an additional figure, which will (among others) highlight some results of this analysis. Figures 1 and 2 in this response show first drafts of such a figure, which is yet to be further improved before inclusion in a revised manuscript.

   In summary, we find a relatively homogeneous dependence of the flowering dates on temperatures (Fig. 1) over all percentiles of the temperature distribution, suggesting that event coincidences on different 'magnitudes of extremes' (i.e. different quantiles being used for event definition) might be expected to show qualitatively similar results. In turn, for precipitation we hardly find any corresponding effect (Fig. 2), underlining the absence of significant influences of precipitation extremes as reported in our discussion paper.

2. *...Yet this results in a multiple testing problem which is not adequately addressed in the paper. [...] A rigorous analysis of this is missing.*

   Since the contents of our discussion paper have already been quite technical, we had not explicitly discussed the multiple testing problem previously. In order to clarify why we have not accounted for this problem, we will add additional text to the manuscript along the following lines:

   First, we generally agree with the reviewer that the lower panels of Figure 3 as well as all the results shown in Figure 4 (of the discussion paper) would require consideration of inflated significance levels. Specifically, the values of meteorological variables

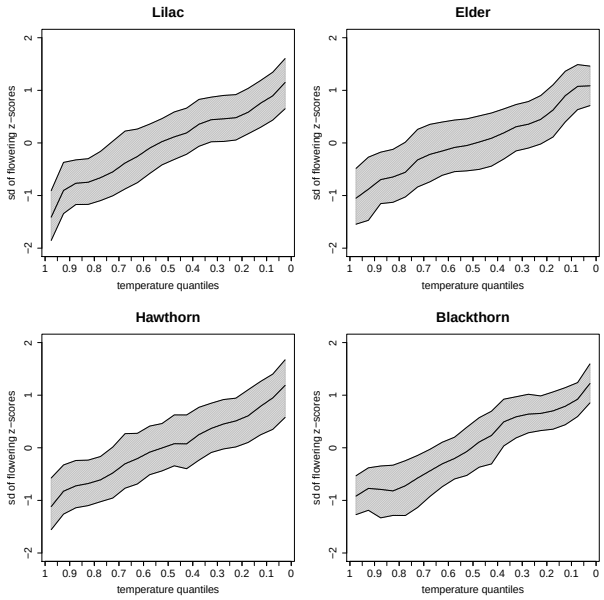

FIG. 1. Mean value (black line) and 25th and 75th percentiles (gray shaded area) of the normalized flowering dates (z-scores) of all phenological stations from 1951-2010 in dependence on the spring mean temperature in the respective year. The temperature quantiles are estimated from all spring temperature values at the respective station.

obtained within our sliding windows are certainly auto-correlated, which is even enhanced by the partial overlap of these windows. Nevertheless, the fraction of stations individually showing a significant number of coincidences is the basic quantity of interest, even if specific statements on statistical significances of this number are avoided. In order to focus the discussion on the practically most relevant information, we will remove the significance thresholds in both figures in the final version of our manuscript, since they could potentially lead to misinterpretations.

Second, the standard approach for dealing with multiple testing problems as present in our case would be a Bonferroni adjustment of the significance level. However, in our opinion such an adjustment is not the appropriate procedure for the present analysis. To see this, we call upon the arguments provided by Perneger (1998):

- *The Bonferroni method is concerned with the general null hypothesis (that all null hypotheses are true simultaneously), which is rarely of interest or use to researchers.* In our paper,

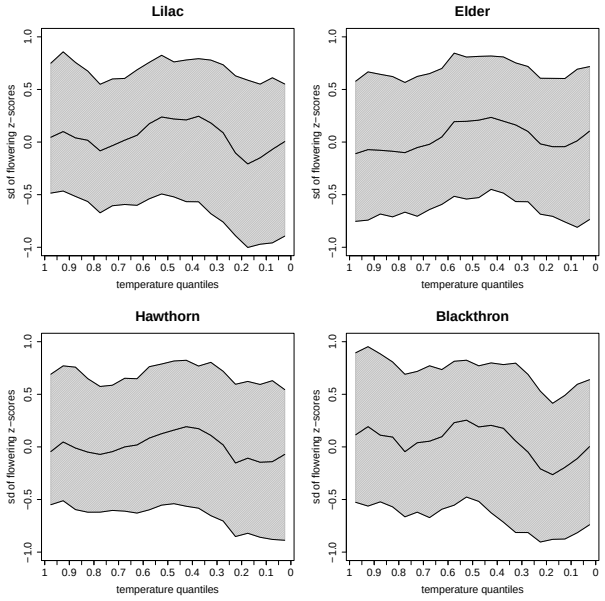

FIG. 2. As Fig. 1 for window-mean daily precipitation.

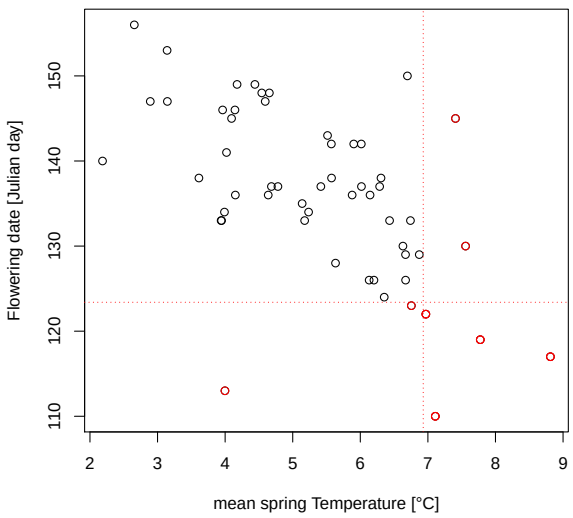

FIG. 3. Scatter plot of flowering dates and mean spring temperatures for one example station. Red dotted lines indicate the corresponding 90th and 10th percentile of the respective empirical distributions.

we do not intend to state that all shrub stands of one species are prone to climate impacts in the same manner, which cannot be expected realistically. In turn, we rather search for a general tendency, which may have a lot of individual exceptions.

- *The main weakness is that the interpretation of a finding depends on the number of other tests performed.* Since the number of phenological stations and, hence, the number of significance tests is larger than 1000 for almost all shrub species, the Bonferroni adjusted α-value would be very close to one. Thus, such an adjustment cannot be of interest for the interpretation of our results, since all (even obvious) interdependencies would be discarded by a test with the accordingly corrected significance levels. Or, put differently:
- *The likelihood of type II errors is also increased, so that truly important differences are deemed non-significant* (Parnegger, 1998).

In summary, we think that the analyses presented in our paper should be left unchanged as far as the multiple testing problem is concerned. However, we will add a brief discussion of this problem to the text in order to help the reader to understand and more adequately interpret the obtained results.

3. *... What is the major advantage of the new approach? What novel conclusions can we draw?*

Since the second reviewer also stated a similar comment (see below), we understand that we probably did not state the conceptual difference between correlation analysis and event coincidence analysis comprehensively enough. It seems that our discussion paper has left the (unintended and misleading) impression that both methods are parallel/interchangeable/competing and necessarily tend to exhibit similar results (or at least results that can be interpreted in a similar manner). However, this is *not* the case. Since correlation analysis includes all observations (i.e., all parts of the distributions of the involved variables), the obtained correlation coefficient is a measure of the common mean behaviour of (or, more precisely, linear relationship between) two variables across their full range of values. It is important to understand that such a strong correlation does not necessarily imply the co-occurrence of extreme values in both records. The latter is only valid if the two variables of interest exhibit a monotonic relationship across all parts of their distributions. As shown in Figs. 1 and 2 of this response, the assumption of such a monotonic relationship is roughly met for temperatures, but hardly for precipitation. However, even in case of temperatures, the flowering response to extreme conditions can be quantitatively different from that to less extreme temperatures (see Figure 3). This is information that can be revealed by event coincidence analysis. In this context, we emphasize that events can also be defined in different ways, e.g., by considering temperature values within a certain quantile range other than the tails of the respective distribution.

Since the scope of our work explicitly related to the

response to extreme conditions, correlation analysis is not the most appropriate tool already for conceptual reasons. Instead, event coincidence analysis has been explicitly designed for the quantification of simultaneities of distinct (extreme) events among paired records. Figure 3 shows one example data set of a single station highlighting the possible mismatch between the timings of extreme temperatures and extreme flowering dates.

Following these considerations, even if the results of event coincidence analysis may (qualitatively) look very similar to results from previous studies using correlation analysis, the findings cannot be interpreted in exactly the same way. The case of qualitatively similar results (as evident for the temperature–flowering relationship) confirms that previous results obtained using correlation analysis are also valid for extreme values (which had not been shown before).

We additionally noticed that Figure 7 of our discussion paper may confuse readers by leading to the wrong impression that both methods are mutually interchangeable and gain equivalent results. Correlation and coincidence analysis show qualitatively similar results *in this specific case*, which is not to be expected in general when comparing these two methods. Therefore, we decided to remove Figure 7 and rather add an additional paragraph to the manuscript, more explicitly emphasizing on the conceptual differences between correlation and event coincidence analysis.

4. *Why can the t-test be used to assess the significance of correlations between binary data (Figure 7). Are the assumption to use the t-test fulfilled? These seems to be questionable.*

   We agree that the assumptions for appropriately performing a t-test are not fulfilled in our case. Note that we provided the corresponding results just as an example of what kind of analysis should *not* been performed in the present context. As stressed above, we will remove the corresponding Figure 7 completely from our manuscript to avoid further confusion on this aspect.

5. *Overall I see some potential in the topic and the methods used, yet many aspects of the analysis are not pursued with the necessary finality and stop halfway.*

   We thank the reviewer for this general encouragement. As stated before, we will add some further aspects to our study by discussing additional results on the impact of extremes depending on the extremes' magnitude along the lines of Figs. 1 and 2 of this response. The latter analysis is not just conducted for temperature and precipitation extremes separately, but also regarding the combined effects of both. A preliminary draft figure of our results on the latter aspect is shown in Fig. 4.

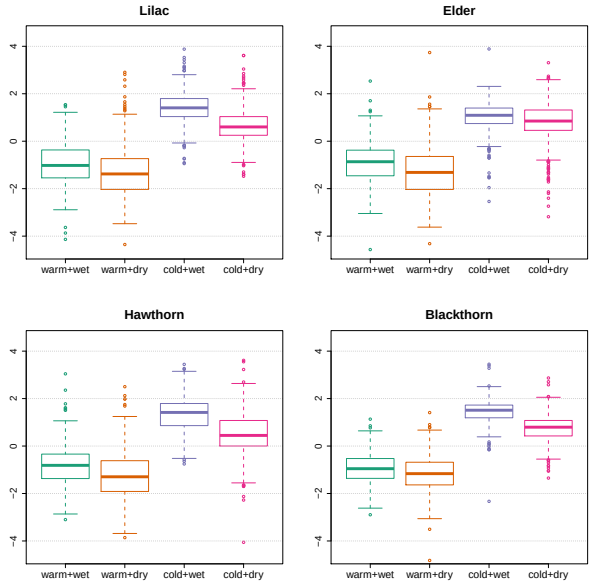

FIG. 4. Box plots of the z-scores of flowering dates of each phenological station illustrating the combined effect of temperature and precipitation extremes on flowering dates of the four considered plant species.

6. *The introduction is rather lengthy and lacks focus.*

   We thank the reviewer for pointing out the corresponding deficiencies of our discussion paper. We fully agree that the introduction should be rewritten in a more concise way. We will follow the according suggestion when preparing our final manuscript, making the specific scopes of the present work better visible from the very beginning.

7. *..., e.g. by using the standardized precipitation index. I encourage the authors to do such an analysis since droughts are among the main causes for strong impacts in ecosystems functioning ...*

   The preprocessing of the precipitation data used in our study (see Section 3.1.2) yields a variable that is (qualitatively) equivalent to the 15-day, 1-month and 2-month SPI, respectively. Therefore, events defined by exceedances of the 90th percentile of the precipitation data correspond to SPI values of moderate dryness (see WMO, 2012). We will emphasize this point in our revised manuscript. Beyond the SPI, we do not think that using further alternative drought indices will be helpful for the present analysis, since the manuscript would clearly loose focus in such a case. However, in order to additionally analyse combined effects of temperature and precipitation extremes, we will include this aspect by providing an additional figure in our final paper

along the results shown in Fig. 4 of this response letter.

8. *Figures 5 and 6: ...If they only represent a negative result (no clustering) this can be stated in words.*

   This comment raises the fundamental question to what extend also negative results should be published. We think that statements on results that are not underlined by figures are of general lower credibility (we already left out all the results for precipitation since no positive results were obtained). In turn, even if we do not see any specific spatial clustering, the shown maps may provide readers (from various scientific backgrounds) with interesting and potentially useful information. As a trade-off, we suggest to remove both figures from the main paper and provide them as a dedicated supplementary information material accompanying our final paper.

In summary, we are confident that our analysis indeed adds some interesting and possibly relevant aspects to existing knowledge on temperature and precipitation effects on plant flowering in the study region. We will highlight these new aspects by a more thorough discussion along the lines outlined above. Our analysis provides a first attempt to statistically disentangle the effects of gradual variability of meteorological conditions versus most extreme (sub-)seasonal conditions, which has not been accounted for in previous studies. We hope that this discussion will fully clarify the rationale of the performed analysis and its possible added value with respect to more classical statistical approaches.

Beyond what can be addressed in the present manuscript, we agree that many of the specific suggestions of the reviewer open important avenues for follow-up studies. Among others, we would like to mention the consideration of different plant types, other phenological phases, additional meteorological variables with known ecophysiological impacts, effects of local conditions (like altitude, slope or soil type), etc. Unfortunately, a systematic investigation of all these aspects is far more than a single manuscript can deliver. Instead, we take the corresponding recommendations as an encouragement for further systematic studies on the aforementioned aspects.

COMMENTS OF REFEREE 2

Following the comments of both reviewers, we realised that the title and introduction of our discussion paper raised too high expectations in comparison to what was actually addressed in the manuscript. This aspect will be clearly improved in our accordingly revised manuscript by rephrasing title and introduction in more precise ways (e.g., emphasizing that only four shrub species and flowering are studied).

The reviewer also stated that one should go more into depth concerning the investigation of precipitation/drought and combined effects of humidity and temperature. This concern will be addressed with new figures (see above), where we not only analyse positive and negative "extremes" in terms of event coincidence analysis, but also illustrate the general impact of meteorological conditions on plant flowering. One corresponding figure will explicitly address combined effects of temperature and precipitation extremes. However, to stay focused, we prefer not to include other (clearly relevant) variables like humidity or soil moisture in our analysis, but perform additional investigations in this direction in a follow-up study.

The reviewer additionally suggests the systematic analysis of delayed effects. Since we already included meteorological data of the complete previous year, it is not fully clear to us which specific additional "delayed effects" have been meant by the reviewer. Effects of extreme conditions on the flowering in the second year (etc.) after the event? Based on our results already shown in the discussion paper, as well as some additional computations on this aspect, we do not expect further similarly strong effects showing up as significant results on this aspect.

Moreover, (similar to the first reviewer) it was stated that *event coincidence analysis is a nice tool, but the paper fails to demonstrate its superiority to a conventional correlation analysis.* We emphasize that in our discussion paper, we never stated that event coincidence analysis is in any way superior to correlation analysis; it is just a conceptually different method with an entirely different meaning and therefore complementary to correlation analysis. Since a similar comment was also made by reviewer 1, we decided to remove the misleading Figure 7 from the paper and instead add a new paragraph explicitly describing the difference between correlation and event coincidence analysis. A more detailed discussion of this aspect has been provided above.

Specific points:

1. *p. 18396 l. 22: "weighted mean interpolation" - weighted with what? The (inverse) geographic distance to the phenological stations?*

   Yes, the inverse geographical distance has been used. We will add the corresponding explanation to the manuscript.

2. *p. 18398 l. 25: "here, $N = M$ by definition" exclude a universe of interesting combination effects (several extreme events in a row might lead to quite different effects than just one, even the latter is bigger) with a few words. Why were the authors forced to this simplification "by definition"?*

   The universe of interesting investigations one could perform is, indeed, extremely large. For this study, as a first attempt to systematically investigate the

impact of meteorological extremes on plant flowering, we decided to use a quite straightforward approach to define extremes by the upper and lower quantiles of the distribution of the respective variable of interest. In further studies we will be glad to adopt these suggestions and further explore this universe of interesting problems.

3. *p. 18403 l. 9/10: where and for what species are these future analyses planned?*

   An interesting field of study would be the Mediterranean region, where previous experimental studies indeed found impacts of droughts on flowering (see the Introduction of the discussion paper for some references). The major challenge for this area will be to obtain appropriate data sets. In addition, we currently extend our reported study by performing similar analyses for different German wildlife plant species and different phenological phases. The aim is to obtain a complete inventory on which plants are to which degree affected by which meteorological variables in which states of their annual development cycle. We plan to publish the results of a corresponding in-depth analysis in a follow-up study. Given also the comments of reviewer 1 and the requirement to stay concise and focused, we prefer not to detail all these future plans in our revised manuscript, since several of the outlined directions of future research are quite obvious, and a large number of issues being highlighted as subjects of future work might leave the wrong impression of immature work.

4. *As association measure between two binary vectors, use the Phi coefficient.*

   We thank the reviewer for this comment. Indeed, the Phi coefficient is a more appropriate measure for comparing binary vectors than a (possibly ill-defined) correlation coefficient. Since we decided to remove Figure 7 from the manuscript, this aspect shall, however, not be further discussed in the revised manuscript.

Regarding the reviewer's impression that *the paper is not finished yet, the really interesting aspects of the relationship between flowering dates and climatic extremes for wildlife plant species are yet to come*, we would like to further refer to our response to the comments of the first reviewer. Indeed, the present manuscript described a pilot study addressing some selected aspects of this much broader problem of interest. We think, that the additionally planed analyses (see Figs. 1-4 of this answer) are a fundamental extension of the discussion paper. However, addressing all questions raised by the reviewers would provide much more results than can be meaningfully described in a single paper. In this spirit, we kindly ask for understanding that extensions of the contents of the discussion paper in addition to what we suggested in this answer (e.g. the impact of several extreme events in a row or taking into account drought indices that include soil moisture) goes beyond what could be done in a revision. All these specific suggestions will be taken into account in follow-up studies.

BIBLIOGRAPHY

Perneger, T.V. (1998): What's wrong with Bonferroni adjustments. In: British Medical Journal, Nr. 316, p. 1236-1238

WMO (2012): Standardized Precipitation Index User Guide. Geneva.