## S1 Introduction

In the following, we first provide additional methodological details on the application of event coincidence analysis. Then, we further elaborate on the differences between event coincidence analysis and correlation analysis based on the consideration of artificial numerical examples as well as the results obtained for the data studied in our main paper. Finally, we provide further results on the spatial distribution of study sites with statistically significant coincidence rates, which may yield initial information on this aspect which could be potentially useful for planning purposes in terms of agricultural and forest management.

## S2 Event coincidence analysis

## S2.1 Analytical significance test

Under the assumption of mutually independent events and, hence, independent exponentially distributed waiting times between subsequent events (corresponding to the null hypothesis of Poisson processes generating the event series), the probability that exactly $K$ coincidences are observed just by chance can be expressed as Donges et al. (2016)
$P(K)=\binom{N}{K}\left[1-\left(1-\frac{1}{T}\right)^{M}\right]^{K} \cdot\left[\left(1-\frac{1}{T}\right)^{M}\right]^{N-K}$
In the present case, $N$ and $M$ denote the numbers of extreme events in temperature/precipitation ( $N$ ) and phenology ( $M$ ) (where we have restricted most of the analyses in our main paper - with the exception of Fig. 7 - to the case of $N=M$ ) and $T$ the length of the time series (number of years of observation). Note that Eq. ( S 1 ) takes the discrete nature of time steps in the phenological records (one year) into account and requires the sparseness of events, a criterion met by the definition of our event thresholds.

Equation (S1) allows defining a simple significance test for the observed number of coincidences ( $K_{\mathrm{obs}}$ ) in two paired event series. For this purpose, we consider pairs of event

$$
\begin{equation*}
\sum_{K \geq K_{\text {obs }}} P(K)<\alpha \tag{2}
\end{equation*}
$$

with $\alpha=0.05$ (0.01) to exhibit a significantly non-random coincidence rate at $5 \%(1 \%)$ confidence level.

We note that the Poissonian assumption is only valid in case of sufficiently rare and temporally uncorrelated events in the two series under study. If the latter are not fulfilled, the expectation value and standard deviation computed from Eq. ( 81 ) are systematically biased. In such cases, we recommend application of Monte Carlo methods for obtaining a constrained resampling estimate of the probability density function of the test statistic $K$ used by event coincidence analysis. For a detailed discussion of these aspects together with different types of event surrogates that can be used for this purpose, we refer to Donges et al. (2016).

## S2.2 Trigger and precursor tests

We emphasize that under general conditions, there are two basic modes to perform event coincidence analysis (Donges et al. (2016)): a "precursor test" (studying the appearance of a preceding climate extreme conditional on that of an extreme flowering date) and a "trigger test" (conditioning the timing of extreme flowering dates on previous extreme climatic events). Since we consider only climatic events at fixed points (windows) in time (instead of allowing for their appearance within a certain period potentially covering several subsequent windows) and have $N=M$, both tests are equivalent in the setting used in this study.

## S3 Event coincidence analysis vs. correlation analysis

In our main paper, we have already discussed the conceptual differences between event coincidence analysis and linear correlation analysis as the statistical approach most commonly used in previous phenological studies. Recall that event coincidence analysis solely
takes the timing of well-defined events in each pair of time series into account (in our case, these events have been defined as the extreme values in the upper and lower tails of the distribution of our variables of interest), whereas correlation analysis uses all explicit values in all parts of the distributions of the variables under study. Accordingly, significant coincidence rates mean "significantly simultaneous events in both time series", while significant correlation coefficients imply "significant co-variability of the two series". Moreover, correlation analysis only captures linear interrelationships between two observables, whereas this restriction is (partially) relieved in the case of event coincidence analysis.

Following these differences, a strong correlation does not necessarily imply the cooccurrence of extreme values (i.e., rare events) in two data sets (and vice versa). The latter would only be valid if the two variables of interest exhibit a monotonic relationship across all parts of the distributions. Such a monotonic relationship between phenological phases and meteorological parameters could be questioned, since the correlation coefficients found in related studies in the past typically ranged between 0.5 and 0.85 . For example, Ahas et al. (2000) reported an $r^{2}$ value between spring temperature and Lilac pollination of 0.52 , i.e., only $52 \%$ of the pollination time variance could be explained by a linear model, whereas almost half of it remained unexplained by this approach. Even in cases where the variance of a phenological phase is much better explained by a linear regression model using a certain meteorological variable as predictor (e.g., $r^{2}=0.75$ for apple pollination and spring temperature, Ahas et al. (2000)), the remaining unexplained variance can still be relevant. Among other possibilities, the extreme values could play an important role for that part of the total covariability that cannot be explained by a linear model.

In this Section, we will provide further evidence that the results of event coincidence analysis cannot be directly inferred from those of correlation-based studies. We first exemplify this claim based on some artificial data sets before proceeding to an inter-comparison between the results of correlation and event coincidence analysis for the data sets under study in our main paper.

## S3.1 Numerical examples

As discussed in our main paper, high (low) correlation coefficients between two time series do not necessarily imply high (low) coincidence rates between the uppermost/lowermost values of these series, especially if the variables of interest exhibit a nonlinear relationship or large noise level. Here, we provide some numerical examples supporting these claims.

## S3.1.1 Critical values of test statistics

To begin with, let us address the different sensitivity and specificity of the resulting statistical tests, which can trigger marked differences between the outcomes of both methods even in the presence of a completely linear relationship between two variables in all parts of their respective distributions. For example, let us suppose a sample size of $T=100$ data points in both variables of interest, and a threshold for defining extremes corresponding to the respective 10th or 90th percentile, i.e., resulting in $N=10$ events in each series. In the case of the linear correlation coefficient $r$, we apply a classical $t$-test with the test statistics

$$
\begin{equation*}
t=\frac{r}{\sqrt{\left(1-r^{2}\right) /(T-2)}}, \tag{3}
\end{equation*}
$$

which asymptotically follows a $t$ distribution with $T-2$ degrees of freedom. At a confidence level of $\alpha=0.05$, this corresponds to a critical value of $r_{\text {crit }}=0.197$ above which an empirical correlation is deemed significant. In turn, under the Poissonian assumption, the probability to find more than 3 (4) coinciding extremes out of 10 within a sample of 100 points is $p(K \geq 3)=0.063(p(K \geq 4)=0.011)$, so that we would consider a critical value of the coincidence rate $\kappa=K / N$ of $\kappa_{\text {crit }}=0.4$. From this, it is evident that a simple scatter plot of correlation coefficient versus coincidence rate might lead to misleading interpretations, since the distributions of both statistics are (beyond exhibiting continuous versus discrete values) not directly inter-comparable.

## S3.1.2 Coincidence rates for nonlinear dependencies

Next, we will discuss a numerical example for which coincidence rates cannot be directly derived from linear correlation values. For this purpose, let us consider one time series $\left\{x_{t}\right\}$ of length $T=100$ being given by simple Gaussian white noise and a second one $\left\{y_{t}\right\}$ generated by the deterministic function $y_{t}=0.3 x_{t}^{3}-b x_{t}$ with a tunable parameter $b$. In Fig. S1. we show the respective medians and $5 \% / 95 \%$ quantiles of Pearson correlation coefficients and coincidence rates (for the uppermost/lowermost 10 values of each sequence) estimated from 100 independent realizations of the noise process $\left\{x_{t}\right\}$. For large $b$, the linear part in the definition of $\left\{y_{t}\right\}$ dominates, and we observe a linear correlation coefficient of $r \rightarrow-1$ and, consequently, coincidence rates between the upper $10 \%$ of $x$-values and the lower (upper) $10 \%$ of $y$-values of $\kappa \rightarrow 1(\kappa \rightarrow 0)$. In turn, for $b \rightarrow 0$, only the cubic term contributes effectively, and for values of $x_{t}$ close to zero (i.e., the majority of values), this cubic term can be roughly approximated by a linear dependence with positive shape, implying that $r \rightarrow 1$ and the coincidence rates behave according to the expectations. The most interesting behavior is found at intermediate values of $b$, where the correlation coefficient between $x$ and $y$ is close to zero, but the coincidence rates differ clearly from zero in many cases. In this regime, the nonlinearity of the function interrelating both variables fully pays out, and the coincidence rates cannot be estimated from the correlation coefficients.


Figure S1. Behavior of Pearson correlation coefficient $r$ (black) and coincidence rates $\kappa$ (green: upper/upper $10 \%$ of values, red: upper/lower $10 \%$ of values) for ensembles of realizations of the numerical example with cubic dependence for different values of the parameter $b$. Thick lines indicate the median, thin lines the $5 \% / 95 \%$ quantiles of the respective values obtained from 100 independent realizations of the noise. The gray lines indicate the corresponding results for $r^{2}$, which coincide remarkably well with those of the coincidence rates in the linearly dominated regime $b \gg 1$.

## S3.2 Correlations and event coincidences between plant flowering and temperature/precipitation

We next illustrate the difference between correlation and event coincidence analysis based upon flowering dates and mean spring temperatures for a single randomly selected study site taken from the data set studied in our main paper. Figure S2 shows that in the considered example, in four out of six cases with the highest spring temperatures, we also observe four out of the top six earliest flowering dates, i.e., we have a significant coincidence rate of $\kappa=4 / 6$. In turn, each of the two variables exhibits two extreme cases with do not coincide with those of the other despite a clearly visible negative correlation.


Figure S2. Scatter plot between the flowering dates and mean spring temperatures for a randomly selected case from the considered phenology data set. Red dots highlight cases with very high mean spring temperatures and/or very early flowering.

In order to further underline the necessity of an event-based statistical approach for studying extreme flowering dates in the study area, we proceed with a corresponding analysis for the whole data set. Figures S3 and S4 show scatter plots between correlation coefficients and coincidence rates for all stations. The calculation of the mean spring temperatures was performed in the same manner as for Figs. 3 and 6 in the main paper. The two figures clearly illustrate that only for roughly half of the stations the correlation coefficient and the coincidence rate are both significant (green circles). In turn, there is a large number of stations where, although the correlation is found statistically significant, the coincidence rate for the extremes is very low and/or not significant (red circles). This effect mainly originates from the much smaller effective sample size utilized by event coincidence analysis which makes the associated statistical tests less powerful. Furthermore, stations with significant coincidence rates (here, commonly above 0.5 ) show a large variability of correlations coefficients ranging from -0.3 to -0.9 , and there are even some stations that show significant coincidence rates without a significant correlation (blue circles). Even though the small event sample size necessarily increases the false positive rate of statistical tests based on event coincidence analysis, the latter finding illustrates again that a high correlation value does not always result in a high coincidence rate for the extremes.


Figure S3. Scatter plot between Spearman's rank-order correlation coefficient and precursor coincidence rate (very warm conditions ( $>90 \%$ ) versus very early flowering ( $<10 \%$ )) of each pair of mean spring temperature and flowering date time series. The time span for the calculation of the mean spring temperature is related to the typical flowering dates of each species: temperature is averaged for the time interval of JD 59-119 for Lilac and Hawthorn, JD 89-149 for Elder and JD 39-99 for Blackthorn. Bold numbers in the lower left corner of each panel give the values of Spearman's rho between rank-order correlation coefficients and coincidence rates for all study sites.


Figure S4. As in Fig. S3 for very cold conditions (<10\%) versus very late flowering (>90\%).

In order to further highlight the importance of the different types of significance tests together with different sample sizes for correlation and event coincidence analysis, Fig. 55 provides a corresponding example showing the fraction of study sites with significant coincidence rates and correlation coefficients between flowering dates and window-mean temperature and precipitation. One clearly recognizes that coincidence rates provide much more conservative indications of interdependencies between the respective variables than correlations.


Figure S5. Fraction of stations with significant coincidence rates (left) and Pearson correlations (right) for Lilac flowering with a window site of 30 days. The significance has been assessed with the binomial test statistic under the assumption of two independent Poissonian processes and a classical $t$-test, respectively.

## S3.3 Correlation analysis based on event data

So far, we have considered correlation analysis and event coincidence analysis as two antagonist methods providing complementary information on the time series under study one being based on the exclusive consideration of information on whether or not a given data point constitutes an extreme situation and the other explicitly utilizing all data points in the time series under study. However, there are methodological alternatives that could provide a reasonable trade-off between both viewpoints. Specifically, we emphasize that the transformation of explicit time series values to event sequences practically corresponds to the generation of a binary time series with values 1 (0) indicating an event (non-event). Even though the direct application of the classical Pearson correlation coefficient (and even more Spearman's rank-order correlation coefficient) to such data is not meaningful, there are powerful alternatives for the analysis of dichotomous variables, such as the $\phi$-coefficient or Cramer's $V$. In the present study, however, the number of events is very small by definition (i.e., the binarized time series include far more zeros than ones), which can be expected to result in values of the latter statistics that are similarly unstable as those of the coincidence rates. In this spirit, we do not expect that the application of these methods provides a significant improvement of our event-based analysis of statistical interdependencies between flowering dates and meteorological conditions. However, this assumption may be challenged and needs further justification by systematic analysis, which we outline as a subject of future research.

## S4 Geographical distributions of significant interdependencies

## S4.1 Latitudinal distribution of significant coincidences for different time windows

## S4.1.1 Warm temperatures and early flowering

Figure 3 of our main paper has already shown the temporal and latitudinal distribution of study sites with significant coincidence rates between high early-spring temperatures and very early flowering for the case of Lilac. Here, we show the corresponding results for the other three considered shrub species in Figs. S6|S8.


Figure S6. Latitudinal distribution (top panels) and total fraction (bottom panels) of stations with significant coincidence rates (red: $\alpha=0.05$, black: $\alpha=0.01$ ) between very early Elder flowering and extremely high window-mean temperatures for three different window sizes. The $x$ axes refer to the starting date of a window. The dashed horizontal lines at $5 \%$ in the lower panels highlight the employed group-significance criterion.


Figure S7. As in Fig. S6for Hawthorn.


19

Figure S8. As in Fig. S6for Blackthorn.

## S4.1.2 Cold temperatures and late flowering

In order to complement the results shown above for warm spring temperatures and early flowering, Figs. S9.S12show the corresponding results for cold temperatures and late flowering for all four shrub species.


Figure S9. As in Fig. S6 for coincidences between cold spring temperatures and late flowering for Lilac.


Figure S10. As in Fig. 59 for Elder.


Figure S11. As in Fig. S9for Hawthorn.


Figure S12. As in Fig. 59 for Blackthorn.

## S4.2 Precipitation effect on flowering dates

In our main paper, we have already described the absence of a statistically significant precipitation effect on the flowering dates of the four considered shrub species. Figure S13 shows the corresponding results for the four possible combinations between extremely wet/dry spring conditions and extremely early/late flowering for all four species, indicating that the fraction of study sites showing a significant coincidence between any pair of extremes hardly ever exceeds the statistical tolerance level of $5 \%$ - the number of false positives to be expected with our test design at an individual confidence level of $\alpha=0.05$.









Figure S13. Fraction of study sites with significant coincidence rates between extremly wet/dry spring conditions and extremely early/late flowering dates for all four shrub species. As before, extremes are defined as events outside the upper/lower decile of the empirical distribution of each variable.

## S4.3 Spatial distribution of significant coincidences with positive temperature extremes

As discussed in our main paper, we have observed significant coincidence rates especially between early flowering and positive temperature extremes. Specifically, the analyses presented there revealed two time intervals of particular interest: late winter / early spring and the previous year's early to mid-autumn. In this section, we further examine the spatial distribution of records with significantly coincident extremes for both time windows.

Lilac


Elder



Blackthorn


Figure S14. Stations with statistically significant coincidence rates between very early flowering and very warm 30-days window-mean temperatures in the time span from 15 March to 30 April (Lilac, Elder and Hawthorn) and 15 January to 15 March (Blackthorn), respectively. Filled black (red) circles mark those stations that show significant coincidences at $\alpha=0.01$ ( $\alpha=0.05$ ) confidence level for at least one window during the aforementioned interval. White circles mark stations that have no significant coincidence for any of the windows.


Figure S15. Stations with statistically significant coincidence rates between very early flowering and very warm 15 -days window-mean temperatures in the period from 1 to 15 September (Lilac, Elder and Hawthorn) and 10 to 20 October (Blackthorn) of the previous year, respectively. Filled black (red) signatures mark those stations, that show significant coincidences at $\alpha=0.01$ ( $\alpha=0.05$ ) confidence level for at least one window during the aforementioned interval. White circles indicate stations that have no significant coincidence for any of the windows.

Figures S14 and S15 show maps with the corresponding results. In order to condensate the potentially large amount of information provided by this analysis, we only plot two maps per plant species representing the two different time intervals. Black (red) signatures mark those stations, which show at least one window with significant coincidences at $\alpha=0.01$ ( $\alpha=0.05$ ) significance level within the time intervals indicated in the respective figure captions. The obtained results allow not only studying the latitudinal distribution of significant coincidences as shown in Fig. 3 of the main paper and Figs. S6. S8 of this Supplementary Material, but also possible patterns or regional clustering of significant results. However, for the 30-days period in spring (Fig. S14, neither a clear pattern nor geographical clusters of stations with significant coincidences are visible.

In contrast to the latter findings, at least the maps for Lilac and Hawthorn in Fig. S15show a weak tendency towards a spatial accumulation of stations with significant coincidences in Northern Germany. In turn, the signatures for Blackthorn concentrate more in the southern part of Germany. However, this observation could also be an artifact of the missing data for most of Northeastern Germany.

## S5 Conclusions

The additional results presented in this Supplementary Material may be used as starting points for further in-depth investigations on various aspects related to both the meteorological drivers of plant flowering as well as methodological aspects. For example, from the visual inspection of the spatial distribution of study sites with statistically significant coincidence rates, it is not obvious if the latter have any statistically relevant underlying pattern. In order to test for the presence of spatial clustering, we outline the application of join-count statistics as a corresponding further research avenue.

## References

Ahas, R., Jaagus, J., and Aasa, A.: The phenological calendar of Estonia and its correlation with mean air temperature, International Journal of Biometeorology, 44, 159-166, 2000.
Donges, J., Schleussner, C.-F., Siegmund, J., and Donner, R.: Coincidence analysis for quantifying statistical interrelationships between event time series, European Physical Journal ST, 225, 471487, 2016.

