GIMA
Geographical Information Management and Applications

Final thesis report

Exploring vegetation seasonality at large scale and determining its uncertainty. A case study with ensemble weather data and the extended spring indices

Rens Vermeltfoort
r.vermeltfoort@students.uu.nl

Raul Zurita-Milla (supervisor)
r.zurita-milla@utwente.nl

Intentionally left blank

Preface and acknowledgements

Before you lies the thesis “Exploring vegetation seasonality at large scale and determining its uncertainty: A case study with ensemble weather data and the extended spring indices”. The thesis was written from January 2020 until August 2020 as part of the graduation of the master program Geographical Information Management & Applications (GIMA). This subject was chosen due to my interest in both Biology and Geography and the connection with the contemporary climatological challenges. Furthermore, this thesis allowed for me to further improve my programming skills in Python. In this thesis, a rewritten Python version of the well renowned extended spring index was implemented in the distributed computing environment Dask. This allowed for relatively fast computations to run the model on all possible weather ensemble members. With the employment of all possible weather ensemble members, the uncertainty of the timing of vegetative events could be assessed.

I would like to thank my supervisor Raul Zurita-Milla for his guidance, support, and valuable feedback on my documents. Furthermore, I would like to thank Gerard van der Schrier for his time to review some of the documents and providing valuable feedback on the climatological aspects of this thesis. I would also like to thank Serkan Girgin for improving parallelism in the Dask implementation and significantly improving the efficiency. Lastly, I would like to thank Theresa Crimmins, Jeff Switzer and Lee Marsh for letting me re-use parts of the code from their extended spring index model implementation.

Thank you for taking your time reading this thesis, I hope you enjoy it.

Rens Vermeltfoort, 14th of August 2020

Abstract

Anthropogenic greenhouse emissions persist to unsettle the global energy balance, causing unprecedented changes in the Earth's climate. Understanding the nature of this change and its impact on human and natural life is a serious scientific challenge. Phenology, the study of cyclic or seasonal natural phenomena, is affected by climate change and can, therefore, be used as an indicator to assess climate change. Climate change can also impact the risk of false springs, the occurrences of late spring freeze after the spring onset. In this study, spatiotemporal patterns of spring onset and false spring risk are examined for Europe with use of the E-OBS dataset. Furthermore, the uncertainty of the predictions is assessed with the employment of the full ensemble of climatological possibilities. To handle the amount of long-term high-resolution gridded datasets on a continental scale on a single device, the modelling is embedded in the distributed computing framework Dask.

This study indicates that spring onset is advancing in Europe, especially in western Europe and mountainous regions. The increase in spring onset was particularly noticeable from 1980 onwards, when global temperatures started to increase rapidly. The change in false spring risk was spatially very heterogeneous, with increases in false spring risk mostly found in the mid-latitudes and decreases in false spring risk mostly found in the higher and lower latitudes. From 1950 until 1979, there was a significant overall increase in false spring risk, whereas the change in false spring from 1980 onwards was negligible and non-significant. The uncertainty of both spring onset and false spring risk is was high in western Europe. The United Kingdom specifically showed high uncertainties in spring onset and false spring risk. The uncertainty in false spring risk was relatively high as compared to the uncertainty in spring onset. The propagation of temperature uncertainty into spring onset uncertainty was highest in western Europe. Furthermore, the mid-latitudes showed higher propagations of uncertainty as compared with the lower and higher latitudes. This study further demonstrates the uniform advancement of spring onset and the spatial heterogeneity of false spring risk change. Furthermore, this study highlights the importance of taking temperature uncertainty into account in phenological modelling, especially when examining false spring risk. The incorporation of temperature uncertainty seems especially relevant in areas with higher uncertainties in phenological outputs, in this case western Europe. Lastly, the Dask implementation proves to be an efficient and relatively uncomplicated solution to the contemporary computational challenges that arise from the ever-increasing volume of geospatial data of this world.

Keywords: Phenology, false spring risk, extended spring indices, uncertainty propagation, big geo-data, Dask.

Table of Contents

List of abbreviations

ANCOVA	Analysis of covariance
CV	Coefficient of variation
DDE2	Growing degree hour accumulation over three days that are 0-2 days prior to the day of calculation
DD57	Growing degree hour accumulation over three days that are 5-7 days prior to the day of calculation
DI	Damage index
FLD	First leaf date, which is a measure of spring onset
GDH	Growing degree hours
HPC	High performance computing
LFD	Last freeze date
MDS0	A counter that starts on January 1st for FLD calculations
MTN	Ensemble mean of minimum temperature
MTX	Ensemble mean of maximum temperature
SI-x	Extended spring index
SYNOP	Season-long cumulative count of high-energy synoptic events
TN	Minimum temperature
TX	Maximum temperature
90% CI	90 percent confidence range

List of figures

List of tables

1. Introduction

1.1 Research context

Anthropogenic greenhouse emissions persist to unsettle the global energy balance, causing unprecedented changes in the Earth's climate. These changes in climate have considerable detrimental effects on human- and natural life. Understanding the nature of climate change and its effects is a serious scientific challenge. Phenology, the study of the timing of life cycle events, provides prominent examples of the impact of climate change on natural life.

Climate change results in mismatches between species. These mismatches result from interacting species from which the regularly repeated phases in their life cycle change at different paces, also known as trophic asynchrony. The impact of such mismatches is dependent on the ecological interaction between the species. When species have a mutually beneficial ecological interaction, the mismatches will have a negative impact on both species. Otherwise, the impact will have negative consequences for one species. Another example of the impact of climate change on phenology can be observed with migration of birds. Migratory birds arrive earlier in their over-wintering grounds due to warming climates.

Vegetation is also affected by climate warming, with advancements in first leaf and flowers in spring and delays in senescence in autumn. Because vegetation is affected by climate change, changes in plant phenology can be used as a biological indicator to study the effect of climate change on natural life. The impact of climate change on phenology is also important economically since vegetation productivity is dependent on the length of the growing season. Moreover, the timing of leafing and flowering is relevant for the agricultural sector and yields may be dependent on the timing of these phenological events. Furthermore, the prevalence of freezes that occur after leaving and flowering may be affected. These subsequent freezes are known as false springs and cause damage to vegetative tissues. False springs can have considerable economic impacts on agricultural yields. In 2012, for instance, a late frost caused damage to fruit trees, resulting in half a billion-dollar losses in Michigan.

Studying the effects of climate change on spring onset and false spring risk is crucial for local decision-makers and environmental management. To study changes in spring onset and false spring risk, models are often employed to estimate historical and future changes in spring onset and false spring risk. These models are often primarily based on temperature accumulation and frequently utilize gridded observational datasets. Regrettably, studies generally do not incorporate uncertainties related to these gridded observational datasets. This thesis will incorporate this temperature uncertainty with the utilization of the full temperature ensembles. These full ensembles consist of equally likely realizations of temperature. The average of the full ensemble is distributed as the 'best guess' dataset and this is the version that is regularly employed for phenological modelling. The extended spring index, a well-renowned model for assessing spring onset and false spring risk, is utilized to make the phenological predictions. Modelling phenology with the full ensemble results in computational challenges since all ensemble members, or equally likely realizations, should be used in the modelling to determine the uncertainty

of phenological outputs. To overcome this computational challenge of handling a full ensemble of long-term high-resolution gridded datasets on a continental scale, the extended spring index is implemented in the distributed computing framework Dask.

1.2 Objectives and research questions

1.2.1 Problem statement and main research objective

Phenological research rarely incorporate temperature uncertainties in their assessments. This may lead to biased results and conclusions regarding spring onset and false springs and potentially inadequate management and decision-making. Incorporating temperature uncertainty requires handling much higher volumes of data, making assessments computationally challenging. This study aims to address the propagation of uncertainty of gridded temperature data to phenological predictions. Temperature ensembles are employed to determine the effect of temperature uncertainty on phenological predictions. With these assessments, the overall timing of spring onset and prevalence of false spring is assessed spatially and temporally. The temporal variability of these phenological predictions is assessed by trend analysis. To make these assessments, a computational solution that enables handling numerous long-term high-resolution gridded datasets (i.e. an ensemble of temperature data) is needed. This main objective translates into the following main research question:

How can the uncertainty of phenological predictions be assessed with utilization of an ensemble of gridded temperature data?

1.2.2 Sub-objectives and research questions

1.2.2.1 Computational solution

The first research objective (SO1) is to implement a distributed computational solution. The implementation of a distributed computational solution is imperative to handle the amount of long-term high-resolution gridded datasets on a continental scale on a single device. This sub-objective comprises the following research questions:

RQ1: How to overcome the computational challenge of handling long-term high-resolution geographical data on a continental-scale?
RQ2: How does the performance the distributed model compare to the performance of the legacy implementation?

1.2.2.2 Uncertainty of spring onset predictions

The second sub-objective (SO2) is to the assess the uncertainty of spring onset predictions with use of individual weather ensemble members. This sub-objective comprises the following research questions:

RQ3: How does the incorporation of temperature uncertainty impact the spring onset trends?
RQ4: How can the propagation of temperature uncertainty into spring onset uncertainty be quantified?

1.2.2.3 Uncertainty of false spring predictions

The third sub-objective (SO3) is to the assess the uncertainty of false spring

predictions with use of individual weather ensemble members. This sub-objective comprises the following research questions:

RQ5: How does the incorporation of temperature uncertainty impact false spring trends?
RQ6: How can the uncertainty of false spring predictions be assessed and quantified?
RQ7: How do the uncertainty assessments vary for different concepts of binary false spring?

1.2.3 Research scope

This study primarily focusses on the assessment of the uncertainty propagation of gridded weather data and the implementation of the extended spring index in a parallelized environment. Implementation of ensemble member data in phenological modeling to assess the uncertainty of phenological predictions is the main novelty that is addressed in this research. The extended spring index (SI-x) model is employed to assess spring onset and false spring risk in Europe.

1.2.4 Research limitations

There are several potentially relevant topics that are excluded in the scope of this research. This section will address the most important limitations in detail.

1.2.4.1 Geographical extent

The geographical extent that is used in this research is limited to Europe (and a part of North Africa). Therefore, the weather data used in this study is also for the extent of Europe. Furthermore, with the spatial aggregation of phenological predictions with the bioclimatic zones for some assessments (Section 3.5) the extent is further limited, as the bioclimatic zones do not extent into Northern Africa. However, the full extent of the bioclimatic zones is not present in the E-OBS dataset either. For instance, everything north of 71.5°N is not included in the E-OBS dataset extent, excluding Spitsbergen from calculations. The original extent of the E-OBS data 25N-71.5N x 25W-45E is modified to 34N-71.5N x 25W-45E to minimize computational load. In this new extend the Canary Islands and part of North Africa are excluded. The remaining number of grid cells is 262.500. This total number of cells include grid cells that represent water bodies.

1.2.4.2 Temporal extent

In this research, the phenological predictions will be made from 1950 onwards since the E-OBS dataset only has temperature data from this year. The last year that is available in the dataset is 2019. Thus, the temporal range is from 1950 until 2019, a total of 70 years.

1.2.4.3 Predictions and forecasts

The products that are derived in this research are based on past weather data. This means that predictions are made based on past conditions. This research will not delve into forecasting of phenology with the inclusion of prospective weather data.

1.2.4.4 Validation of phenological model

In this research, the extended spring index is used to assess phenology and false springs (Section 3.2). The validity of the extended spring index has been tested extensively by studies in the past (Schwartz et al., 2006, 2013; Schwartz et al., 2000).

Therefore, the validity of the model is accepted a priori, and no comparisons will be made between the products made in this research and phenological ground observations.

1.2.4.5 Model predictors

The extended spring index uses latitude and temperature variables as its input. Furthermore, several intermediary variables are created with the extended spring index. The relative influence of these primary and intermediary variables on the phenological products may differ temporally and spatially. For instance, the relative influence of short term growing degree hours and season-long cumulative count of high-energy synoptic events on phenological products were found to vary geographically (Zhu et al., 2019). This research, however, will not assess the relative influence of these primary and intermediary variables.

1.2.4.6 Modulating effects of natural climate variability

Some studies incorporate the effect of natural climate variability caused by large-scale climate modes, such as oceanic oscillations, on phenological trends (Labe et al., 2017; McCabe et al., Palecki, 2006). The effect of such large-scale climate modes on phenological trends will not be explicitly studied in this research.

1.2.4.7 Working with ensemble data

The E-OBS ensemble consists of a 100-member ensemble (Haylock et al., 2008). Each individual ensemble member is viewed as a separate realization of climate, which is independent of the other ensemble members. To approximate the uncertainty of computations the separate realizations of reality are employed, which leads to 100 computations per grid cell per year.

2. Theoretical framework

The existing literature concerning phenology, false springs, uncertainty in weather data, and computational challenges regarding handling big data are reviewed in this second chapter.

2.1 Phenology and spring onset

Phenology is defined as “the study of the timing of recurrent biological events, the causes of their timing with regard to biotic and abiotic forces, and the interrelation among phases of the same or different species” by the International Biological Program (IBP) (Lieth, 1974). The timing of the biological events is influenced by habitat factors, such as nutrient availability or elevation, and by seasonal and interannual fluctuations in climate. The word phenology is derived from the Greek word *phaino*, which means to appear or to show. Phenological events include, but are not limited to, the emergence of leaves and flowers, the timing of migration of animals, hatching of birds, the timing of leaf coloring in fall, and hibernation of animals. Phenology has a long history that can be dated back thousands of years ago when people realized that documentation of recurrent phenomena could be useful for making decisions in agriculture. Over the last centuries, phenology as a field itself has been viewed with some indifference. In the past, phenology suffered the status of being performed by amateur naturalists, but not as an innovative science. The last decades, scientific interest in phenology has much increased due to its importance in monitoring climate change (Menzel et al., 1999; Piao et al., 2019; Schwartz et al., 2006; Schwartz et al., 2000). Phenological events are very sensitive to temperature changes driven by weather and climate. Hence, phenology can be used as an indicator of long-term biological impacts of climate change on the timing of plant and animal life cycle events (Peñuelas et al., 2002; Schwartz, 2003).

Plant phenology encompasses all vegetative life cycle events. Examples of plant phenology include first leaf of the year, first flower of the year, first fruit of the year, senescence of leaves in fall, and leaf drop in the fall. It is demonstrated that spring onset is one of the most reliable biological indicators of climate change (Schwartz et al., 2006). Due to the high sensitivity of spring onset to climate variability, spring onset is especially useful for studying the effect of climate change on vegetation (Cayan et al., 2001; Schwartz et al., 2006). Spring onset consists of the first leaf and flowering of plants after winter dormancy. Due to climate change, spring onset is likely to occur earlier (Schwartz et al., 2006, 2013; Zhu et al., 2019). Changes in spring plant phenology are extensively observed for in situ observations in Europe, North America and Eastern Asia (Piao et al., 2019). From these in situ observations there is consensus that spring is mostly advancing in these regions (Dragoni et al., 2011; Fu et al., 2015). This advancement is further shown from phenological modeling with climate data in the Northern Hemisphere (figure 1) (Schwartz et al., 2006). This advancement of the start of spring is especially noticeable from 1980 onwards (Dai et al., 2019). There is variability in the amount of advancement in the different studies, species, and regions (Piao et al., 2019). In Northern America spring

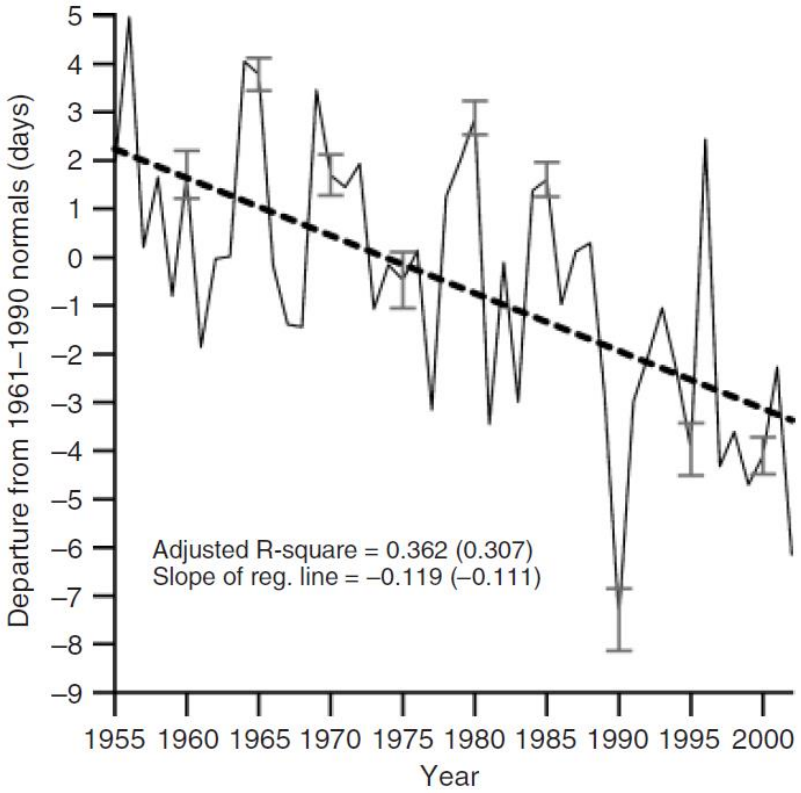


Figure 1. Advancement of spring onset in the Northern Hemisphere approximated by phenological modeling. The standard error is shown at 5-year intervals with the error bars. The linear regression trend is shown with the heavy black dashed line (Schwartz et al., 2006)

seems to be advancing slower as compared to spring advancement in Europe. Satellite observations further confirm the advancement of spring throughout the northern hemisphere (Stöckli et al., 2004; Zhou et al., 2001). However, while in situ and satellite observations show overall spring advancement from 1980s onwards, there is evidence from studies looking at satellite observations that the advancement of spring onset has been reduced or even reversed from the 2000s (Jeong et al., 2011; Piao et al., 2019). This stagnation of spring advancement may be due to the warming hiatus period, which is a slowdown in global warming from 1998 until 2013 (Karl et al., 2015).

There are many factors influencing spring onset. These factors include humidity, precipitation, soil moisture, soil temperature, light regime, photoperiod, nutrient availability, and air temperature (Dai et al., 2019). From these factors, air temperature is the most influential when it comes to spring onset (Dai et al., 2019; Schwartz, 2003). There is a direct relation between warmer years and earlier springs. Likewise, cooler years have later springs as a result (Menzel et al., 2006). The relationship between spring onset and temperature, however, is mostly nonlinear (Fu et al., 2015). The photoperiod is another important driver in plant phenology. The photoperiod is the time of a day in which an organism receives illumination. Photoperiod co-regulates spring onset through its interaction with temperature.

There are various models that predict spring onset based on several variables. These models generally include a component that accumulates temperature over time. Such models include the Spring Warming Model (Sarvas, 1974) and the Thermal Time

Model (Cannell et al., 1983). Growing degree days (GDD), a measure of daily temperature accumulation, is a concept that is often employed in such models. Growing degree hours (GDH) and growing degree minutes (GDM) show higher accuracy in heat summation than GDD (Gu, 2016).

2.2 False springs

The onset of spring has advanced in the recent decades due to a rise in global temperatures. This overall advancement of spring onset may result in longer growing seasons and vegetation productivity (Menzel et al., 1999; Peñuelas et al., 2001), which in turn may increase carbon uptake by vegetation and diminish climate change (Dragoni et al., 2011). On the other hand, advancement of spring onset may result in mismatches between timing of phenological events and animal species dependent on these events (Kellermann et al., 2015; Schweiger et al., 2008). Moreover, earlier spring onset in combination with greater temperature fluctuations in some regions could lead to damage to vegetation due to freezes that occur after spring onset (Ault et al., 2013; Gu et al., 2008). The spring in 2012 in North America is exemplary of the impact early spring onset may have on agricultural yields. In this year, the spring onset was very early due to abnormally high temperatures in the start of the year. However, due to subsequent freezing there was much damage to plants, resulting in major financial losses in the agricultural sector (Ault et al., 2013). A start of spring, which is then interrupted by a damaging spring frost, is called a false spring. The term 'false spring' has always be accompanied with a great deal of ambiguity. Vulnerability to frost damage, for instance, varies across tissues and also

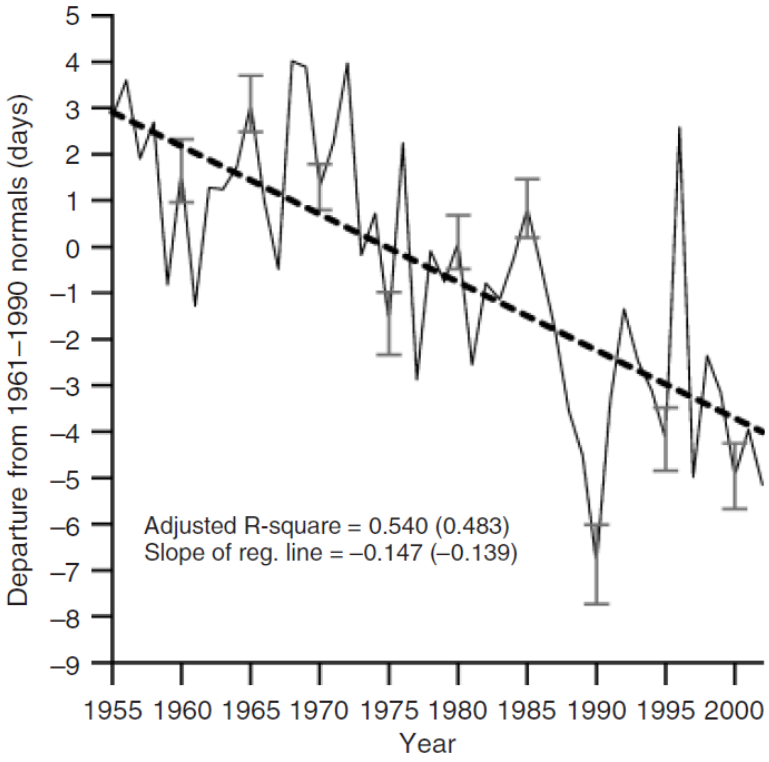


Figure 2. Advancement of last spring freeze date across the Northern Hemisphere. The standard error is shown at 5-year intervals with the error bars. The linear regression trend is shown with the heavy black dashed line (Schwartz et al., 2006)

seasonally with plant development (Chamberlain et al., 2019). Some tissues in general are less vulnerable to frost damage than other plant tissues. Flower tissue and fruit tissue are very sensitive to frost and are easily damaged by false springs (CaraDonna et al., 2016). Contrarily, wood, bark and leaf tissues show less sensitivity to frost and in general are better capable of surviving false springs (Charrier et al, 2011; Strimbeck et al., 2015). Development of cold hardiness (or freezing tolerance) in vegetative tissues is the primary cause for seasonal variability in frost tolerance. Cold hardiness consists of physiological mechanisms that allow vegetation to resist cold temperature better (Strimbeck et al., 2015). Besides the variability in different seasons, species and tissues, the actual timing of the false spring is relevant for the vulnerability of the plant tissues. Often, the later the frost occurs after initial spring onset, the more damage occurs in the plant tissues (Allstadt et al., 2015; Peterson et al., 2014). Therefore, when modeling false spring risk, often earlier and later false springs are distinguished (Peterson et al., 2014; Zhu et al., 2019).

Like spring onset, last spring freeze dates are advancing across the Northern Hemisphere over the last decades (figure 2) (Schwartz et al., 2006). However, there is no scientific consensus as to whether the risk of false spring is increasing or decreasing due to climate change. Some studies find that cold weather diminishes faster than that spring onset is advancing, which would result in decreased risk of false springs (Peterson et al., 2014; Schwartz et al., 2006). Contrarily, other studies show an overall increased risk of false spring due to a faster advancement of spring onset as compared to the diminishment of cold weather (Zhu et al., 2019). Most studies agree that the risk of false springs will vary locally, where some locations will have an increased risk and others a decreased risk of false springs (Allstadt et al., 2015; Zhu et al., 2019). False spring risk can be associated with circulations in the atmosphere. In western Europe, these atmospheric circulations bring cold and clear air from the north, which allows radiation to escape at night and reduces temperature significantly. A climate change signal in these atmospheric circulations has not been detected. Therefore, this phenomenon is likely to persist in warmer climates (Belmecheri et al., 2017). This could be an indication that false spring risk will increase in regions that are influenced by these atmospheric circulations, including western Europe.

2.3 Uncertainty of gridded temperature data

Many different models have been employed to extrapolate in-situ observations of spring onset to unvisited areas (Czernecki et al., 2018; Mehdipoor et al., 2020). These models are designed to predict spring onset in various locations. Temperature is often the primary driver in spring onset and false spring risk (Dai et al., 2019; Schwartz, 2003). Therefore, to make spatially continuous phenological prediction, spatially continuous temperature data is required. Gridded weather data is frequently used as an input in phenological models (Izquierdo-Verdiguier et al., 2018; Schwartz et al., 2013; Wu et al., 2016). Gridded weather data that is derived from interpolating data from weather stations are a representation of reality and is inherently subject to uncertainty (Cornes et al., 2018; Scully, 2010). There are different sources of uncertainty that are inherent in these gridded observational datasets. Zumwald et al. (2020) distinguishes three general sources of uncertainty in observational temperature datasets (figure 3). The first type of uncertainty arises during the generation of the dataset and involves how an environmental parameter is measured (1a) and how the measurement outcome is further processed (1b). The

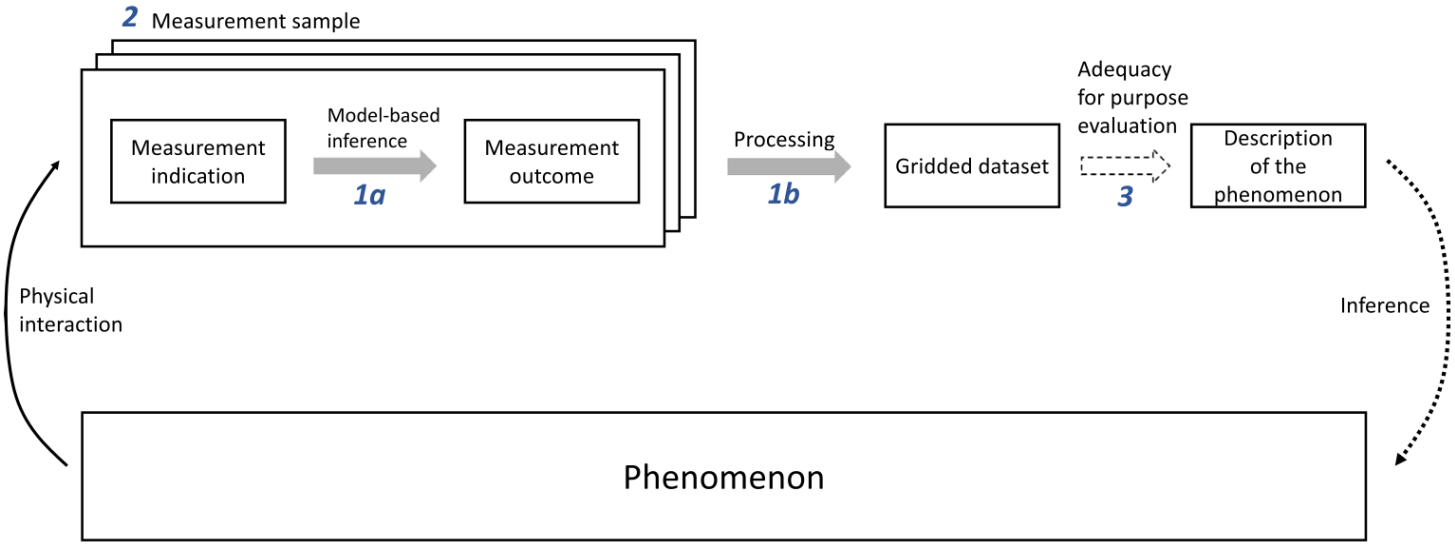


Figure 3. Graphical representation of the three sources of uncertainty in observational gridded datasets. Uncertainty related to the accuracy of a measurement (1a) and the subsequent processing (1b); uncertainty related to the degree of representativeness of a sample (2); and uncertainty related to the adequacy of certain properties for a specific cause (3). In reality, these sources are more complex and mutually dependent (Zumwald, 2020).

second source of uncertainty is uncertainty that deals with where and when a measurement is conducted, and it ultimately concerns the representativeness of the measurement sample (2). The third source of uncertainty concerns the uncertainty that arises from the adequacy of abstract properties for a certain purpose (3). These abstract properties include the spatial and temporal resolution of a gridded dataset, the metric of a dataset, and the unit in which a dataset expresses its value. The first two types of uncertainty are representational uncertainties, whereas the third type represents a non-representational source of uncertainty (Zumwald et al., 2020).

Ensembles may be employed to account for the different types of uncertainty of a dataset. Factors that induce uncertainty can be methodically varied to create the dataset ensembles. These variations include the parameter values, the model structures, the measurement sample, and the dataset properties (Zumwald et al., 2020). From these different possible variations, four types of dataset ensembles can be distinguished. From varying the parameters, and applying different plausible parameter values, *parametric ensembles* are constructed. A *structural ensemble* of datasets helps assess structural uncertainty, which arise through underdetermination of modeling approaches in measurement device construction or processing of measurements. *Resampling-based ensembles* assess uncertainty through passing subsamples of measurements into the processing procedure. Generally, resampling helps address biases but is not employed to correct them. Lastly, the *property ensembles* are created by varying abstract properties of the ensemble, such as resolution or metric that is used (Zumwald et al., 2020).

The E-OBS dataset is a daily gridded temperature and precipitation observational dataset that is created for the range of Europe (Haylock et al., 2008). The dataset is a parametric ensemble, in which parameter values are varied to create a 100-member

ensemble (Cornes et al., 2018). These parameters can be used in different ways to interpolate the weather station data. Decisions must be made regarding features that affect the interpolating algorithm and the resulting gridded dataset. Such features include the search radius for inclusion of influencing stations and the impact of co-variates such as latitude, elevation, and distance to the nearest water body (Cornes et al., 2018). This results in an ensemble of possible gridded datasets resulting from differing interpolations. The mean of the ensemble of gridded dataset is labeled as the best approximation and distributed as a dataset. These best approximation datasets are frequently used in phenological modelling (Izquierdo-Verdiguier et al., 2018; Wu et al., 2016). Uncertainty of phenological model outputs deriving from gridded temperature uncertainty is rarely assessed. For this purpose, the spread of temperature could be integrated when modeling phenology. The spread is the difference between the lower and upper percentiles over the ensemble, which indicates a correspondent uncertainty range (Cornes et al., 2018). However, a more thorough approach to assess uncertainty propagation would be to incorporate full weather ensembles in the modelling process.

2.4 Big data & computation

Ecology and other Earth sciences increasingly must deal with big data in the recent years and decades. Throughout the domain of ecology, volumes of databases are increasing rapidly (Farley et al., 2018). Furthermore, the variety, complexity, and heterogeneity of data is increasing, and one of the main challenges of ecology is to structure and order the abundance of data (LaDeau et al., 2017). Due to a rapid increase in citizen-based science efforts the credibility of ecological data can be at stake (Farley et al., 2018). Lastly, the increased rate of data generation may require high-velocity analytical solutions and iterative modeling (Dietze et al., 2018). Phenology, as part of ecology, is no exception in this case, and technologies are advancing to keep up with contemporary challenges. For instance, the consistency of volunteered phenological observations are checked (Mehdipoor et al, 2018), phenological modeling and satellite-based vegetation metrics are coupled with machine learning (Czernecki et al., 2018), and iterative plant phenology forecasting is being automated (Taylor et al., 2020).

Phenological modeling often involves dealing with gridded observational datasets (Section 2.3). Through technological advancement, the resolution of these gridded observational datasets is increasing over time (Farley et al., 2018). Multiple solutions have been employed to work with long-term high-resolution gridded datasets on a continental scale in phenological modeling. These solutions often integrate scalability, which is the ability of a computer application to function for increasing volumes of data. While it is possible to acquire hardware that can load and process high volumes of weather data, it is less costly and more efficient to perform downscaling and iteratively execute computations on smaller portions of weather data. Various frameworks that employ scalability have been used for phenological modelling, including Apache Spark (Zurita-Milla et al., 2019), Google Earth Engine (Izquierdo-Verdiguier et al., 2018), and Dask (Taylor et al., 2020).

Pangeo is a new open-source community driven platform that enables scalability to meet current and future challenges of big data (Abernathey et al., 2017). The core mission of Pangeo is to develop a cooperative environment in which open-source analysis

tools for Earth sciences can be developed and sustained. The core technologies of the Pangeo platform are the Python packages Dask (Dask Development Team, 2016) and Xarray (Hoyer et al., 2017). Dask is a flexible parallel computing library with dynamic task scheduling possibilities. Xarray is provides conventional data structures (e.g. arrays, datasets) with user-friendly meta-data tracking, visualization, and indexing. Both Dask and Xarray are integrated with netCDF datasets, a standard file format for large or complex geodata which is commonly used in the field of Earth science. The Pangeo platform connects end users to a high-performance computing (HPC) system through Jupiter Notebook on a conventional internet browser. The user can then perform data analysis on Xarray. Dask is employed to schedule computations across computer nodes, allowing parallel reading of data from the storage system as necessary (figure 4) (Abernathey et al., 2017). In many aspects, Pangeo is similar to other big-data libraries, such as Hadoop or Apache Spark. These libraries also enable analysis of bigdata on HPC computing and facilitate parallelism. The primary advantage of the Pangeo platform over existing tools is its versatility (Abernathey et al., 2017, Xu et al., 2019). Hadoop and Apache Spark are primarily oriented towards tabular data structures and cannot effortlessly ingest large multidimensional numeric arrays, whereas Pangeo provides more efficient handling of large multidimensional arrays. This makes the Pangeo platform especially appropriate for Earth sciences such as ocean, atmosphere, and climate sciences, since large multidimensional arrays are relatively common in those sciences.

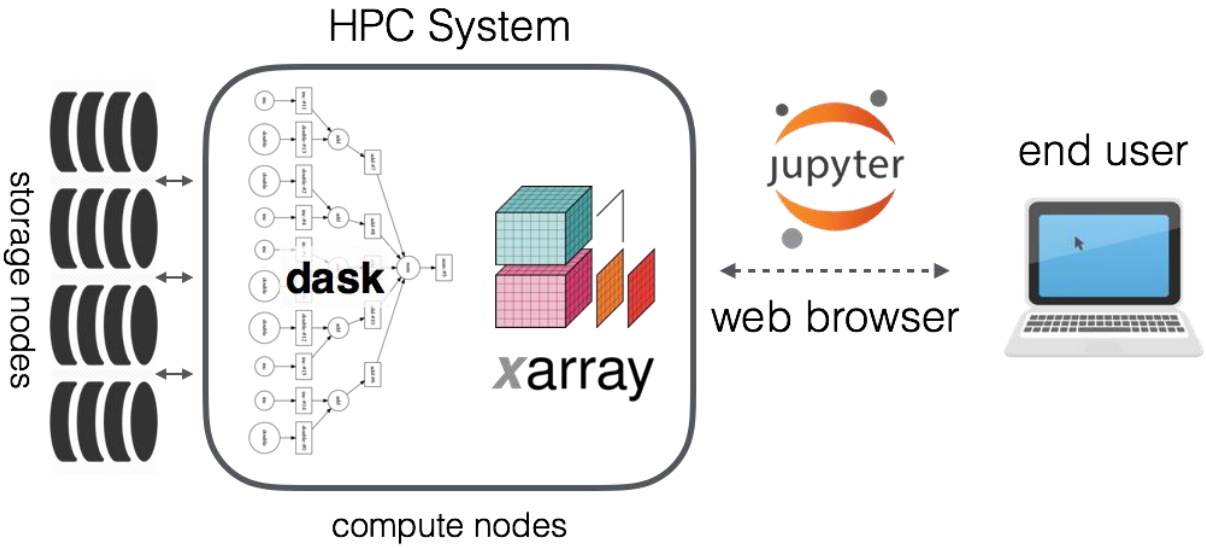


Figure 4. Schematic overview of the Pangeo elements and their interrelations. The end user is connected to the HPC system through Jupyter Notebook on a regular internet browser. The user can perform data analysis in the Xarray environment on the HPC system. Dask regulates task scheduling and facilitates parallel reading of data.

3. Materials and methods

In this third chapter, the materials and methods that are used to answer the research questions are explained. Firstly, the temperature data that is used in the computation is presented. Secondly, the extended spring index is explained. Thirdly, the software that is employed is reported and the computation and adjustments to the original model are explained. Fourthly, the false spring assessments are made clear. Lastly, the assessments of uncertainty of both spring onset and false spring risk are explained.

3.1 Temperature data

The E-OBS dataset is used in this research. This dataset is a product of the European Climate Assessment and Dataset (ECA&D) project and covers the whole of Europe as well as Northern Africa and Turkey (Haylock et al., 2008). ECA&D combines daily observations at meteorological stations and creates datasets consisting of daily temperature, precipitation, radiation, and sea-level pressure. For this research, we use the daily maximum (TX) and minimum (TN) temperature with a spatial resolution of 0.1 degrees (version 21.0e). The full ensemble will be used in this study to account for the uncertainty of gridded temperature data. This full ensemble for TX and TN consists of a 100-member ensemble. The ensemble means of minimum temperature (MTN) and the ensemble means of maximum temperature (MTX) are employed to compare with the outputs from this research (Section 3.4). Figure 5 shows the geographical extent of the dataset that is used in this study. The figure depicts the spatial distribution of the maximum temperature on an arbitrary day (1 April 1999).

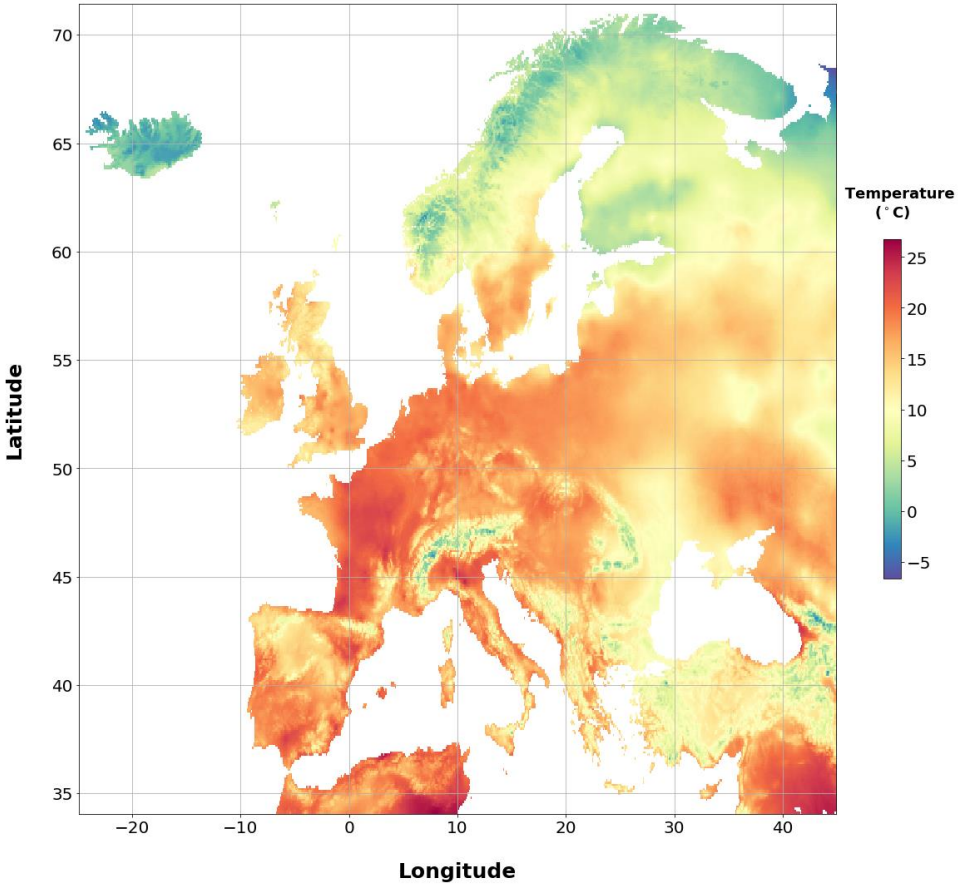


Figure 5. The geographical extent of the E-OBS dataset that is used in this study. The figure shows the spatial distribution of the maximum temperature in Celsius on an arbitrary day (1 April 1999). The blue color indicates lower temperatures and the red color correspond with higher maximum temperatures

3.2 Extended spring index

The extended spring indices (SI-x) are utilized to calculate spring onset. The SI-x models predict the first leaf date (FLD) and first bloom date for three indicator species, namely one lilac (*Syringa chinesis "Red Rothomagensis"*) and two honeysuckle clones (*Lonerica tatarica "Arnold Red"* and *L. korolkowii "Zabeli"*). Daily minimum and maximum temperature and latitudinal information of the sites are the inputs for the model (Schwartz et al., 2013). From the daily minimum and maximum temperature, the growing degree hours (GDH) can be calculated, which is the foundation of the SI-x models. The exact workings of the model for the three different species are shown in the equations below:

$$Eq(1) \qquad DDE2 * 0.201 + DD57 * 0.153 + SYNOP * 3.306 + MDS0 * 13.878 \geq 1000$$

$$Eq(2) \qquad DD57 * 0.248 + SYNOP * 4.266 + MDS0 * 20.899 \geq 1000$$

$$Eq(3) \qquad DDE2 * 0.266 + SYNOP * 2.802 + MDS0 * 21.433 \geq 1000$$

The equations show the model and its predictors for the lilac (Eq (1)), the Arnold Red honeysuckle (Eq(2)), and the Zabeli honeysuckle (Eq(3)). In these equations, DDE2 is the GDH accumulation (base temperature of 0.6 °C) over three days that were 0-2 days prior to the current date of calculation. DD57 is the GDH accumulation (base temperature of 0.6 °C) over the three days that were 5-7 days prior to the current date of calculation. SYNOP is the number of high-energy synoptic events that have occurred since MDS0, which is a counter that starts on January 1st for FLD calculations. FLD equals the date at which the cumulative weighted addition of DDE2, DD57, and SYNOP calculations is equal to or surpasses 1000, an arbitrary value that is used to calculate the weights of Eq(1-3).

In this study, the first bioindicator FLD will be used to determine spring onset and the subsequent first bloom date will not be considered. The FLD is especially relevant for the scope of this research as it the bioindicator that is used in false spring risk predictions. Moreover, FLD has been capture overall ecosystem green in general, including spring onset of shrubs, grasses, and fruit trees (Allstadt et al., 2015; Schwartz et al., 2006, 2013). The employment of different phenology models and temperature datasets to calibrate the models prevent accurate comparisons of the results among different regions and species. Utilizing a widely used and well-parameterized model facilitates adequate comparisons and reduces uncertainty (Zhu et al., 2019). In this research, spring onset or leaf out is characterized as the average of the FLD of the three indicator species (Ault et al., 2015a). Since only the FLD will be considered in this thesis the model will henceforth be called the extended spring index and not the extended spring indices.

3.3 Software and computation

3.3.1 Software

Python is the main software that is used in this research. Python is an open-source high-level programming language that is often employed in the scientific community (Downey, 2015). Due to the high level of abstraction, Python language it is relatively easy to use and, therefore, especially convenient for people with no comprehensive knowledge in computer science. Another major benefit of Python is the extensive availability of libraries. A library in a programming language is a collection of precompiled methods

which can be used in a program. The Xarray[1] (version 0.14.1), Dask[2] (version 2.1.0) and NumPy[3] (version 1.16.4) libraries are employed for the SI-x model runs. The subsequent computation section (version 3.6.2) will elaborate on the implementation of these libraries in the computation. The matplotlib[4] (version 3.1.0) library is employed for visualization purposes and the SciPy[5] (version 1.2.1) library for analysis.

3.3.2 Computation

The extended spring indices were originally written in the Fortran programming language. Ault et al. (2015b) translated the original Fortran code to MATLAB and developed supporting documentation to guide model users (Ault et al., 2015b). MATLAB, however, is a proprietary language that is not freely available for everyone. Therefore, in this research a translated Python version of the original model will be used (figure 8A1). As mentioned in the previous sub-section, another notable asset of Python is the availability of many libraries.

```
File "C:\Users\Rens\Anaconda3\lib\site-packages\xarray\coding\variables.py", line 72, in __array__
  return self.func(self.array)

File "C:\Users\Rens\Anaconda3\lib\site-packages\xarray\coding\variables.py", line 217, in _scale_offset_decoding
  data = np.array(data, dtype=dtype, copy=True)

File "C:\Users\Rens\Anaconda3\lib\site-packages\xarray\coding\variables.py", line 72, in __array__
  return self.func(self.array)

File "C:\Users\Rens\Anaconda3\lib\site-packages\xarray\coding\variables.py", line 142, in _apply_mask
  return np.where(condition, decoded_fill_value, data)

MemoryError
```

Figure 6. When the 0.1-degree TN data is loaded in python, a memory error occurs.

Working with the high-resolution 0.1-degree temperature data poses a significant challenge. When a temperature grid is loaded on a single device with 16 GB RAM, a memory error occurs (figure 6). The data is too large to handle on a single device. Most python libraries, such as NumPy, are not originally designed handle data that does not fit the memory. A distributed computational framework must be employed to handle the many computation on high resolution. The Python library Dask provides the tools to enable parallel computing. Dask parallelizes many libraries in the Python ecosystem, including Pandas and NumPy. Dask allows the libraries to scale either on a single machine with multi-core memory parallelism, or on large distributed clusters (on a cloud, for instance). Dask is integrated with existing libraries to enable easy transitions from traditional single machine workflows to parallel and distributed computing without the need to learn new frameworks or rewriting all the code. In this research, Dask will be implemented locally on a single machine (figure 8A2). This is the default setup for Dask and is relatively easy to configure. Even though there is no cloud computing, the model will be run in parallel on a

[1] http://xarray.pydata.org/en/stable/
[2] https://dask.org/
[3] https://numpy.org/
[4] https://matplotlib.org/
[5] https://www.scipy.org/scipylib/index.html

single machine, resolving the memory issue. The computation will be parallelized on different batches, where each worker will do the computation of only that batch. The batches are split along the latitudes. The optimal batch size/number of batches to run the model on will be determined and subsequently the data was run in that specific batch size.

3.3.3 Code adjustments

The code mostly has the same structure as the MATLAB implementation (Ault et al., 2015b). However, the bloom calculations in the original code are not used in this research. The influence of the bloom calculation on the performance of the model was tested to ascertain an optimal model performance for the scope of this research. In a test run with a smaller spatial and temporal extent, the speed was tested for calculating the FLD with the original code (including bloom calculations) and an implementation from which all bloom calculation was excluded. The implementation excluding bloom calculations performed over two times as fast as the original implementation. This is a significant difference, especially since the model must be run on larger spatial and temporal extents of a 100-member ensemble. Therefore, the redundant bloom calculations are eliminated from the original code in this implementation to minimize computational load. Furthermore, vectorization has been applied to parts of the model to minimize the number of loops, as loops generally are computationally expensive. The optimization is based on code provided by Crimmins et al., (2017)[6]. These optimizations resulted in year-by-year calculations, as compared to MATLAB all-in-one calculations.

3.3.4 Model verification and benchmarking

The new Dask implementation will be verified by comparing the model outputs to the model outputs of the MATLAB implementation. The results from both implementations are checked for three sub-regions and three years to ascertain similar behavior spatially and temporally. The ensemble number that is used in the computation is randomized. The three sub-regions that are used for this comparison are at different latitudes to get spatially diverse comparisons. The sub-areas that are used in the computation are 3-degree square regions, consisting of 900 grid cells. Figure 7 shows the locations of the sub-areas that are employed in comparing the FLD output from both

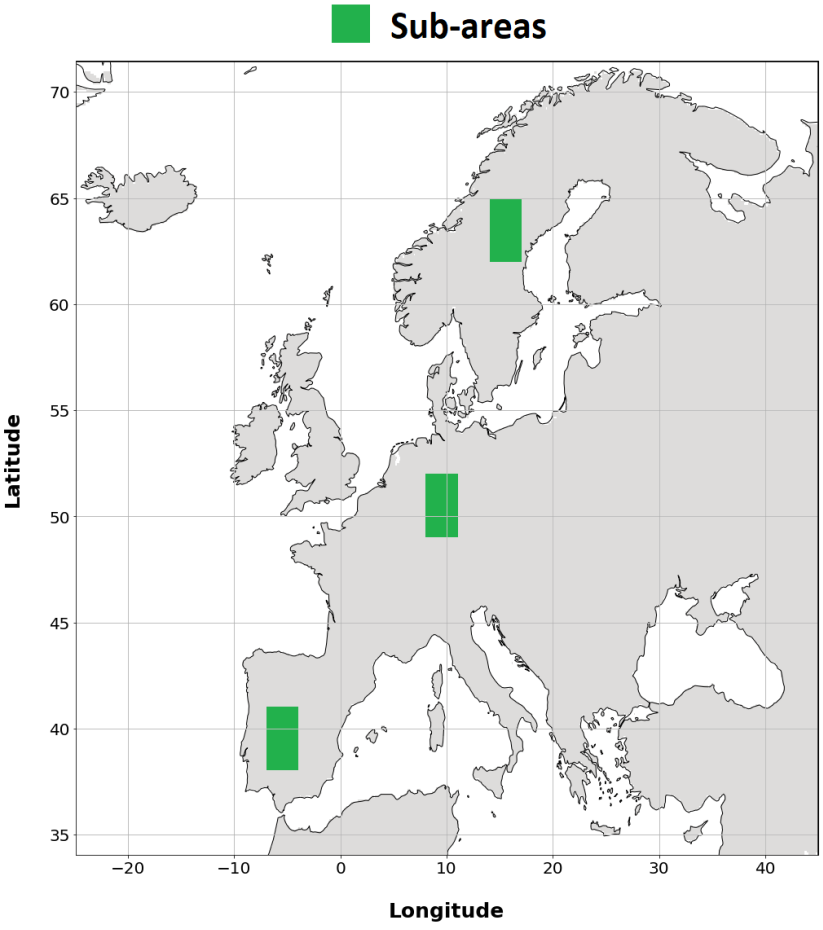


Figure 7. The sub-areas that are used for the MATLAB implementation and Dask implementation comparison.

[6] https://github.com/usa-npn/gridded_models

implementations. This comparison is done by Pearson correlation of the outcomes for both implementations. If the Pearson correlation is 1, then the implementation that is used here is identical to the MATLAB implementation (figure 8B). Furthermore, the RMSE is calculated to determine any errors between the FLD output of both implementations. Besides the verification of the Dask implementation, the different SI-x models that may be employed to calculate FLD and LFD are compared in their computation time. The implementations that are run and compared are the MATLAB implementation, the direct Python translation of the MATLAB code, the optimized Python model, and the parallelized optimized model in the Dask environment. Since the full ensemble of one year does not fit into memory, the first 10 ensemble members are used in the calculation.

3.4 False spring risk

The SI-x model produces derivative products, besides the FLD and first bloom date calculations, that enable false spring risk calculations. These derivative products include calculation of the last freeze date (LFD) and the damage index (DI), which is the difference in days between the FLD and the LFD. The last day prior to the 1st of June with a temperature below -2.2 °C qualifies as the LFD (Schwartz et al., 2006). The DI has been extensively used to assess false spring risk (Allstadt et al., 2015; Izquierdo-Verdiguier et al., 2018). From the DI, binary definitions of false spring are employed to assess false spring risk (Allstadt et al., 2015; Peterson et al., 2014; Zhu et al., 2019). These binary false springs are either false spring or no false spring, depending on the LFD in comparison to the FLD. However, even for the binary definition of false spring risk, there are varying ideas of what constitutes a binary false spring (Allstadt et al., 2015; Chamberlain et al., 2019). This ambiguity is due to the fact that vegetation is more vulnerable in later phenological stages (Augspurger, 2013). Following Peterson et al. (2014), this study employs a 0, 7, 10, 14-day lag between FLD and LFD to calculate the binary false springs. The added value of the DI is that it registers early LFDs in combination with late FLDs as positive values, whereas this information is lost with the binary false springs. Contrarily, the different binary false springs provide insight into the uncertainty of earlier and later false springs specifically.

3.5 Uncertainty assessment

For the determination of the uncertainty of phenological predictions, the 100-member ensemble weather data is utilized. The SI-x model is run on all 100 members of the ensemble, resulting in 100 FLDs and false spring date calculations per grid cell per year (figure 8B). The calculations will be carried out for the study period 1950-2018. Grid cells may miss FLD values for some year. For instance, in very cold regions or mountainous areas where the cumulative weighted addition of DDE2, DD57, and SYNOP may not surpass 1000 for some years. Furthermore, if fewer than four stations are found in a distance from 500 km from the grid cell, the value is set to missing (Cornes et al., 2018). Therefore, grid cells with sufficient FLD values in the study period are considered for trend analysis. Following Schwartz et al. (2006) and Ault et al. (2015a), at least 80 percent of the years should have valid FLDs. This means that at least 56 years out of the 70-year reference period should have valid FLD values for the trend analysis. Missing years are removed from the trend analysis.

Figure 8. Flowchart of the methods of this research. The first steps of this research entail implementing the original MATLAB model in a python Dask environment, running the model on all ensemble data, and verifying the new model (A, B, C). Thereafter, temporal trends can be made of the FLD and DI outputs, and the outputs from the full ensemble may be compared to the outputs from the ensemble mean model runs (D, E, F, G, H, I). Besides the trend analysis, the uncertainty of outputs is examined. The uncertainty of spring onset outputs is calculated, and the propagation of uncertainty is determined (J, K, L, M, N). Likewise, the uncertainty is calculated for the false springs (O, P, Q, R).

3.5.1 Spring onset uncertainty

To assess the impact of working with the full ensemble on the FLD trends, the trends created from the averaged FLD from the full ensemble is compared to the trends from the ensemble mean. The average FLD from the full ensemble is not necessarily the same as the FLD derived from the ensemble mean (MTN and MTX), since the SI-x model captures non-linearity in the accumulation of temperature. Therefore, the FLDs of the full ensemble are averaged for each grid cell/year (figure 8D). From these averaged FLDs, trends are generated with the parametric linear regression test (Helsel et al., 1992) (figure 8E). The parametric linear regression is usually employed for phenological trend analysis and is based on ordinary least squares estimation (Bock et al., 2014; Dai et al., 2014). However, the parametric linear regression is sensitive to outliers in the temperature data. Therefore, besides the linear regression, the more robust Theil-Sen estimator is employed to analyze trends. The Theil-Sen estimator is a method that fits a line to the sample points by choosing the median of the slopes of all lines through pairs of two-dimensional sample points (Theil, 1992). The advantages of the Theil-Sen estimator compared to linear regression include simplified computation, robustness to outliers, better capability of testing assumptions, analytical estimates of confidence intervals, and requires less information regarding measurement errors (Fernandes et al., 2005; Wang et al., 2018). The trends of the full ensemble FLDs are compared to the trends created with the FLDs that are acquired from running the model with MTN and MTX (figure 8F/8G). Besides the slopes, the statistical significance of the different trends is computed for each grid cell (figure 8H). The linear regression and Theil-Sen Estimator are estimated with functions from the Python package SciPy. The trends that are derived from the ensemble mean and the full ensemble are compared with analysis of covariance (ANCOVA). (figure 8I). The change in FLD value over time is the independent variable. The year of observation is used as a covariate. The effect of the year of observation is of secondary interest and controls the main effect of the independent variable. The analysis of covariance is performed with functions of the statistical Python package Pingouin.

The uncertainty of the FLD calculations is measured as the 90 percent uncertainty range (90% CI) of FLD and the coefficient of variation (CV) (figure 8J). The 90% CI of is also used in quantifying the propagation of temperature uncertainty. Larger widths of the 90% CI and higher CV values indicate higher uncertainty of FLD model output. To illustrate this, if the 90% CI for FLDs for a particular grid cell for a specific year is from 90-120 (Julian dates), then the range which is considered in the calculation of the overall uncertainty of FLD is 30 (the width of the 90% CI). The relative spread of 90% CI and the CV values are averaged over the study period to summarize uncertainty of FLD predictions over all grid cells (figure 8K). The change of uncertainty over the study period through trend analysis is not assessed since the uncertainty would change primarily due to changes in weather station density, and not due to changes in climate.

To quantify the propagation of uncertainty of FLD from the gridded temperature data, the uncertainty range of both TN and TX is compared to the confidence interval of the associated FLD. As the GDH is calculated from hourly temperatures, which are interpolated from daily TN and TX, the average confidence interval of the average temperature is a good approximation of temperature uncertainty. The spread of the average temperature of the ensemble is employed as a measure of temperature uncertainty (figure 8L). The spread is calculated as the difference between the 5^{th} and 95^{th} percentiles calculated from the 100 members at each grid cell. The ensemble spread provides a measure indicate of the 90% uncertainty range. The spread is averaged from the 1^{st} of January until the average FLD of the associated grid cell to compress the temperature uncertainty to a single value (figure 8M). The propagation of temperature uncertainty into FLD uncertainty can be quantified by comparing the compressed temperature uncertainty value with the 90% CI of FLDs (figure 8N). Figure 9 shows the concept of uncertainty propagation quantification graphically. The individual dots

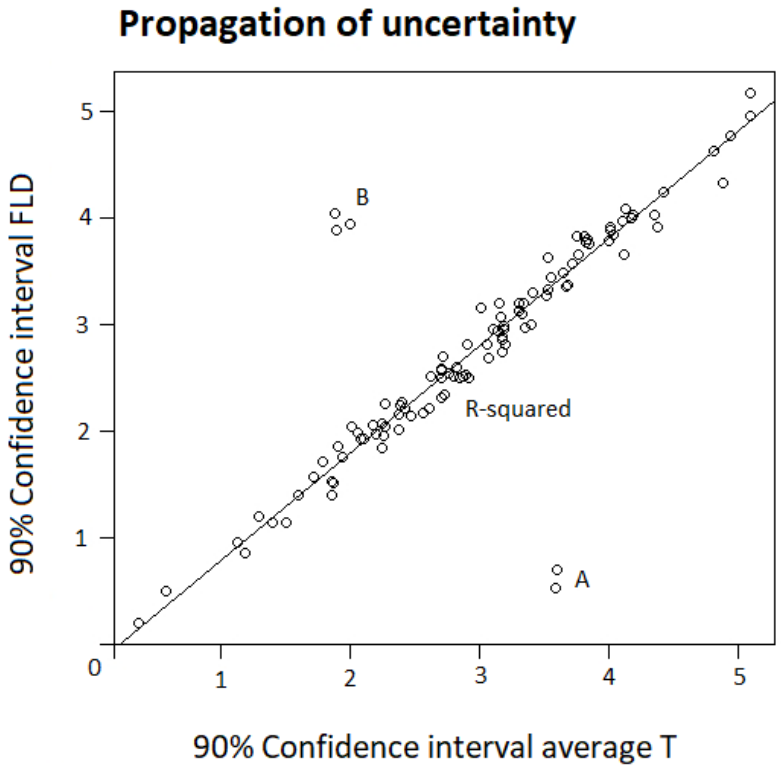


Figure 9. Illustration of the quantification of temperature uncertainty propagation to FLD model output. The average width of 90% uncertainty range of average temperature for the days up until the FLD is displayed on the x-axis. The width of the 90% uncertainty range of FLDs is displayed on the y-axis. The R-squared is a measure of the strength of the relationship between average temperature uncertainty and FLD uncertainty, where high R-squared values represent a strong relationship. The individual dots represent the different grid cells. The propagation of uncertainty temperature into FLD uncertainty is low for the cluster of grid cells at A, whereas the propagation of uncertainty is high for the grid cells at B.

represent the different grid cells. The propagation of uncertainty temperature into FLD uncertainty is low for the cluster of grid cells at A, whereas the propagation of uncertainty is high for the grid cells at B. As there will be no trend analysis for the uncertainty propagation quantification, the quantification of uncertainty will be averaged over the years in the study period.

3.5.2 False spring uncertainty

To assess the impact of working with the full ensemble on the false spring trends, the trends created from the averaged DI from the full ensemble is compared to the DI trends from the ensemble mean. To do this, the DIs of the full ensemble are averaged for each grid cell/year (figure 8D). From these averaged DIs, trends are generated with linear regression and the Theil-Sen estimator (figure 8E). The trends of the full ensemble DIs are compared to the trends created with the DIs that are acquired from running the model with MTN and MTX (figure 8F/8G). The statistical significance of the trends is computed for each grid cell (figure 8H) and the differences between the trends are assessed through the ANCOVA test (figure 8I).

The uncertainty of the false spring is assessed differently for the DI index and for the binary false springs (figure 8O). The uncertainty of the DI is assessed with the 90% CI that derives from the full ensemble outputs (figure 8P). The CV is not considered for DI uncertainty, as the mean DI may be close to 0, which will result in spurious CV values. The binary false spring data, however, is ultimately categorical data, and the uncertainty is in the likelihood of

the occurrence of a false spring. If a binary false spring occurs a value 1 is assigned and if a false spring does not occur a value of 0 is assigned. The value that is received for the 100-member ensemble will be a probability value (p-value) between 0 and 1. The uncertainty of the outcome of the 100-member ensemble is maximum at 0.5, since the occurrence of a false spring is as likely as it is unlikely. The uncertainty of a false spring occurrence, therefore, is best quantified with a function that assigns high uncertainty values for p-values that approximate 0.5, and low uncertainty values for values that are closer to either 1 or 0. The uncertainty H(p) from the binary Shannon entropy function is an appropriate function that assigns uncertainty approximations in this fashion (figure 8Q) (figure 10) (Shannon, 1948). Lastly, the probability values and uncertainty values from the different binary false springs can be compared (figure 8R).

There are two main inputs responsible for the uncertainty of false spring predictions, namely the uncertainty of FLD and the uncertainty of the LFD. As the false spring calculations are the result of the subtraction LFD from FLD quantifying the uncertainty in the same way as with the FLD predictions would be inconsequential as the quantification would be the direct result of both uncertainties. Therefore, the influence of both input uncertainties on the uncertainty of false springs is measured as the 90% CI of the FLD and the LFD.

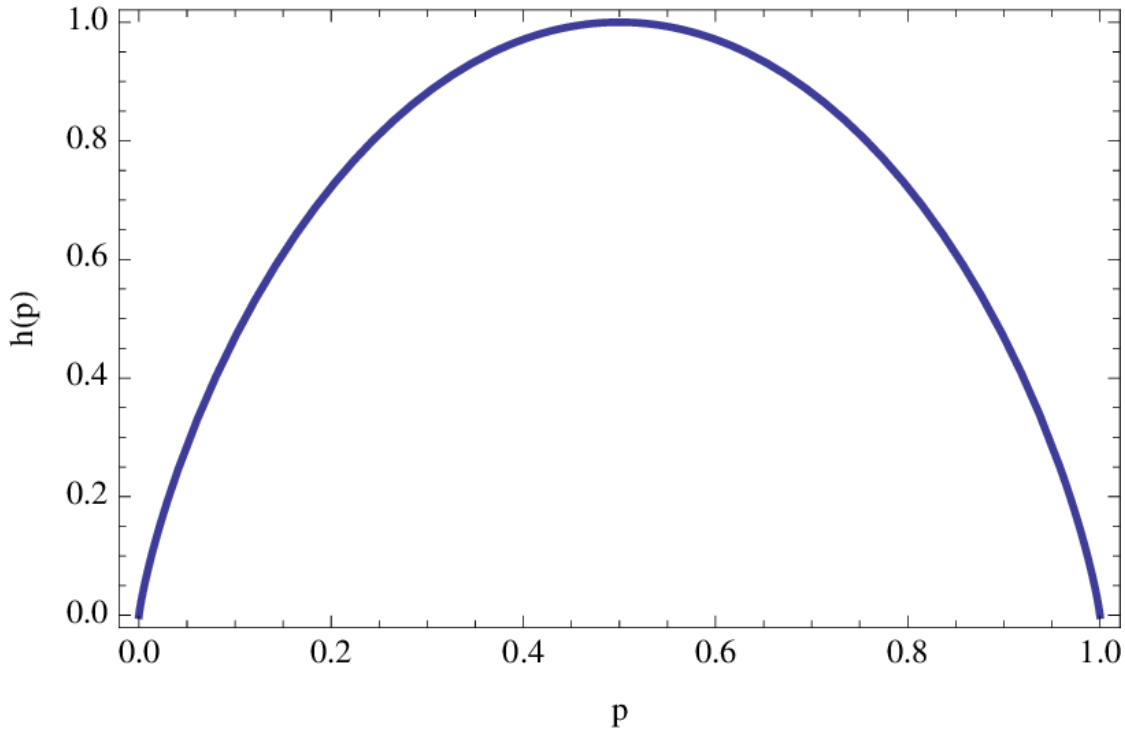


Figure 10. Binary Shannon entropy function which calculates the uncertainty (H(p)) as a function of the probability (p-value).

.

4. Results

The results from this research are presented in this fourth chapter. Firstly, the performance of the Dask implementation is discussed. Secondly, the spring onset results, with the corresponding trends, are discussed and presented. Thirdly, the false spring risk outputs and trend analysis are considered. Lastly, the uncertainty assessments of both spring onset and false spring risk are conveyed.

4.1 Performance evaluation

The optimal batch size proved to be 75 latitudes per batch, resulting in 5 batches per ensemble computation. For all the different spatial and temporal extents, the Pearson's correlation between the FLD obtained with the Dask implementation and the FLD obtained with the MATLAB implementation was 1. Moreover, the RMSE values that were calculated between the different implementations and different temporal and spatial extents were all 0. The computation time for 1 ensemble members for a single year was a little less than 3 minutes. This makes the computation of the 100-member ensemble for one year approximately 4.75 hours. The total computation time for the 262.500 pixels for 70 years and 100 ensemble members was approximately 330 hours on a single device. The computation time for the ensemble mean for 70 years was 3.5 hours in the Dask implementation. The computation time for the different model implementations is shown in table 1. Initially, the Dask computation is much faster than the direct Python translation with an almost 24 times faster computation time. After making the optimization adjustments, the MATLAB implementation was still faster than the optimized Python implementation, with a computation time of a little less than twice as fast. The Dask implementation is the fastest implementation, with a computation time of 37.6% of the computation time of the MATLAB implementation.

Implementation	*Computation time (s)*
MATLAB	3189
Python translation	76212
Python optimized	5852
Dask implementation	1201

Table 1. The computation time of the different possible implementations in seconds. The computation times are based on the computation of 10 ensemble members. The computation times of 4 different implementations are compared, namely, the MATLAB implementation, the direct translated Python implementation, the optimized Python implementation, and the Dask implementation.

4.2 First leaf dates

The first leaf date values were calculated for all the cells in the spatial extent of the dataset and for all ensemble members and the years and in the temporal scope. The spatial and temporal average of the FLD is on the Julian date 101, or the 11th of April. Figure 11 shows the averaged FLD output for all ensemble members and all the years. Earlier FLD values correspond with greener colors, whereas yellow colors correspond with relatively late FLD values. The variation of FLD output is reasonably broad, with the earliest FLD on the 25th of January and the latest FLD on the 22nd of July. In general, the FLD is later in northern Europe as compared to southern Europe. Furthermore, the FLD seems to be later in eastern Europe as compared to western Europe for the same latitudes. Even though the FLD is best explained by the South-to-North gradient and West-to-East gradient, there are numeral exceptions to this

generalization. For instance, there are several 'cold spots' that have later FLD values than the grid cells from the same latitudes. Prominent cold spots include the Alps, the Pyrenees, and the Caucasus Mountains. Moreover, lower mountainous regions such as the Spanish Meseta, the Italian Apennines, the Scandinavian mountains, and the Scottish Highlands are also distinguishable due to later FLD values as compared to their immediate surroundings. Even though the pattern shown in figure 11 resembles the pattern of maximum temperature for a single day (figure 5), the spatial patterns are fundamentally dissimilar. The straightforward difference is that figure 5 shows the temperature for a single day, whereas figure 11 depicts the average spring onset for multiple years and ensemble members. A more intricate distinction between the establishment of the patterns is that the SI-x model captures non-linearity in the accumulation of temperature and is therefore more than the simple accumulation of growing degree days (Schwartz et al., 1988; Wu et al., 2016). The non-linear relationship between temperature and FLD emerges from the inclusion of high-energy synoptic events that serve as capstones to spring onset, the so-called the 'capstone effect' (Schwartz et al., 1988).

The variability of FLD over the years is shown in figure 12. The variability is lowest in the south of Spain and Italy and the largest part of North Africa. Furthermore, the variability is relatively low in Scandinavia and the East European Plains in Western Russia. The variability is highest in Western Europe, with high standard deviations in the United Kingdom, the Netherlands, and Denmark.

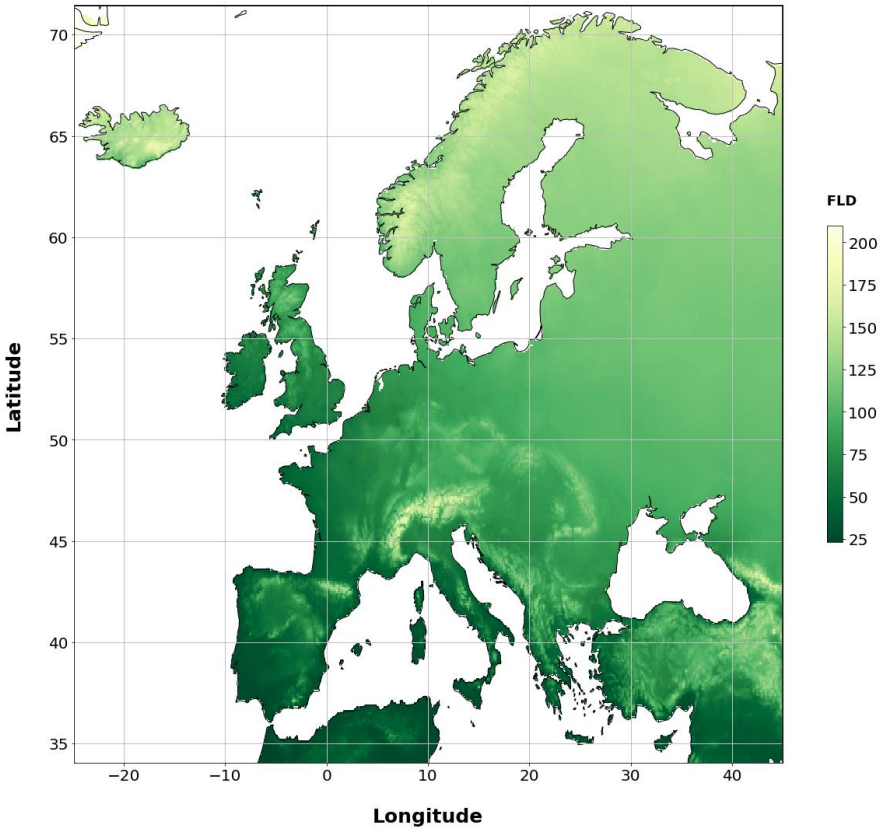


Figure 11. The spatial distribution of the average FLD for all ensemble members and years. Earlier FLD values correspond with greener colors, whereas yellow colors correspond with relatively late FLD values.

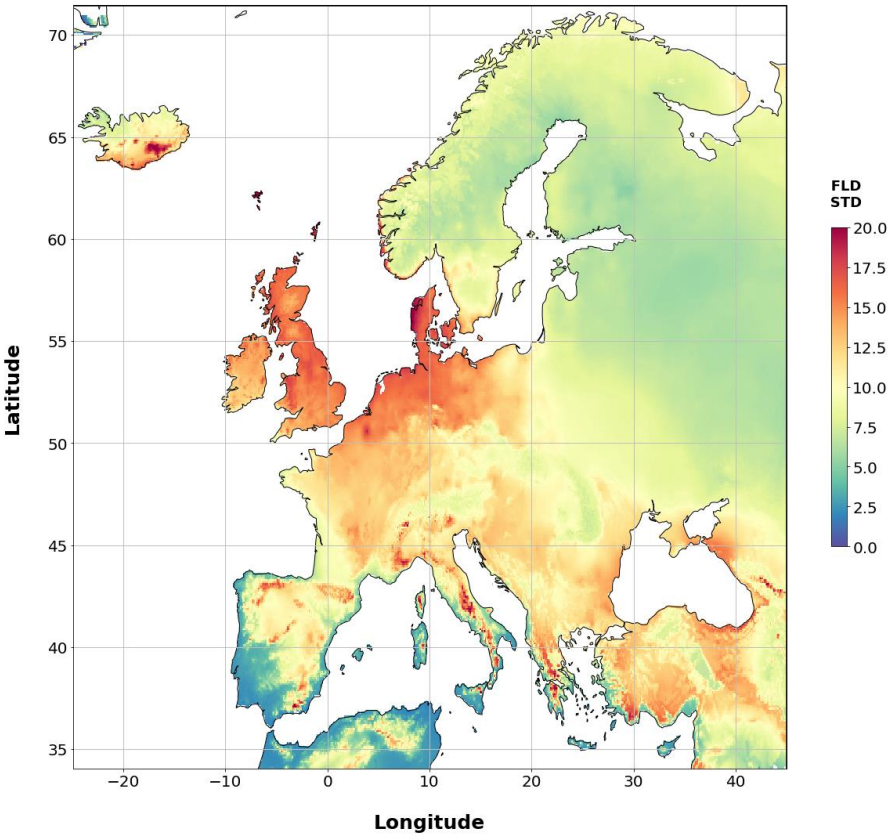


Figure 12. The across-year variation of the average FLD of the 100-member ensemble. The blue color corresponds with low across-year variabilities, whereas the red color corresponds with high across-year variabilities.

Moreover, regionally higher variabilities exist, predominantly in mountainous regions, such as the Alps, the Apennines, the Caucasus, the Pyrenees, and the Spanish Meseta.

4.2.1 FLD trends

The spring onset advances on average with 2 days per decade for all grid cells for the full temporal range (table 2). However, the slope of the trends is low from 1950 until 1979 (0.2 days per decade) and higher from 1980 until 2019 (3.1 days per decade). Furthermore, the trends are significant for the full temporal range and from 1980 until 2019 (p-value of 4.02E-11 and 2.83E-06, respectively), and non-significant for the temporal range from 1950 until 1979 (p-value of 0.78). This means that there the warming temperatures from 1980s onwards induced an advancement of spring onset. Figure 13 shows the average FLD for all grid cells per year and the corresponding trends for the different temporal ranges. A sharp decrease in FLD is noticeable from 1980 onwards.

Period	*Slope*	*P-value*	*Std. err.*
1950-1979	-0.02	0.78	0.08
1980-2019	-0.31	2.83E-06	0.06
1950-2019	-0.2	4.02E-11	0.03

Table 2. FLD slope values for three different temporal ranges, namely 1950-1979, 1980-2019, and 1950-2019. All slopes are negative and the slopes for the entire temporal range and for 1980-2019 are significant.

Both the Theil-Sen slopes (figure 14A) and the linear regression slopes (Appendix A) indicate that the FLD is decreasing for most grid cells. The downwards slope is highest for countries in western Europe, especially for the United Kingdom and Denmark. Furthermore, mountainous regions, such as the Caucasus and the Alps, appear to have sharp decreases in FLD values. The significance of the trends is high in almost all regions with a negative trend, approximating a p-value of 0 (figure 14B). The sparse regions with positive trend values and slope values approximating 0 uniformly show relatively high p-values, indicating that there are no regions with significant increases of FLD over the study period.

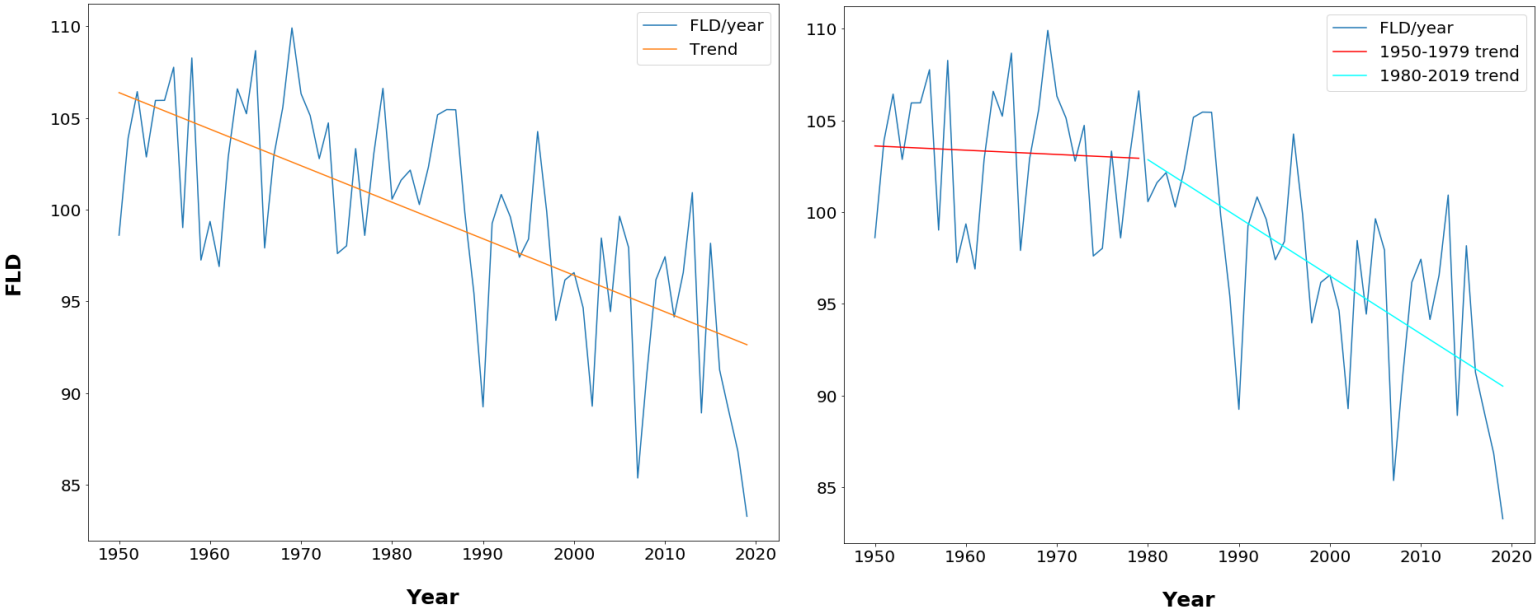


Figure 13. The average FLD for all grid cells per year and the corresponding trends for the different temporal ranges. The left figure shows the trend for the full temporal range with the orange line. The right figure shows the trend for the 1950-1979 in red and the trend for the 1980-2019 period in cyan. A sharp decrease in FLD is noticeable from 1980 onwards.

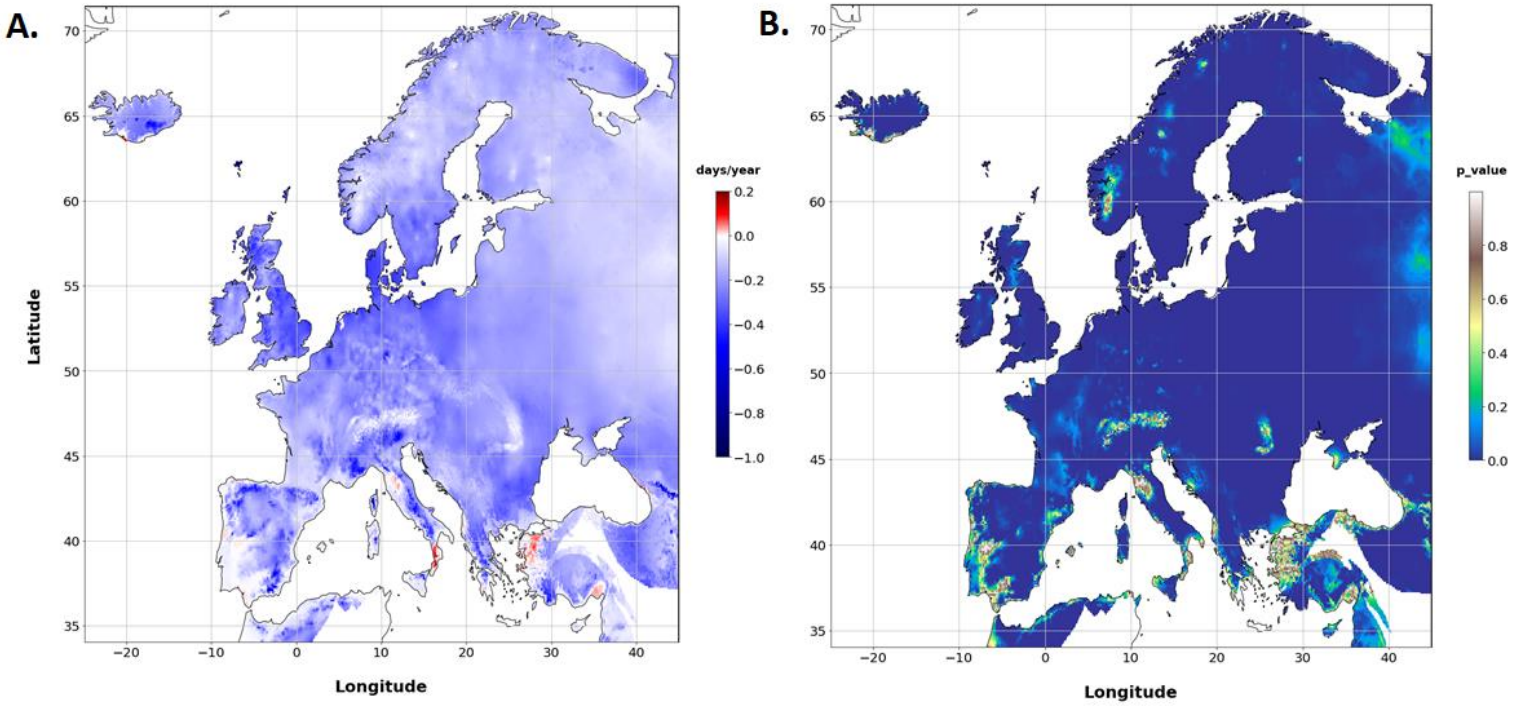


Figure 14. (A) The FLD Theil-Sen slope values. Most grid cells have negative slope values, indicating a decrease in FLD values over time. (B) The statistical significance of the slope values. Most grid cells with negative slope values have p-values approximating 0, whereas positive slope values and slope values of approximately 0 show relatively high p-values and low significance.

4.2.2 Impact of full ensemble on FLD trends

To assess the impact of employing the full temperature ensembles on the FLD outputs and trend assessment, the model was run on the ensemble means. The difference between the average FLD of the full ensemble and the FLD resulting from the ensemble mean was calculated for each year and the absolute differences were averaged over the study period (figure 15). Regions with very early FLD values showed the least difference between the two methods, approximating an average difference of 0 days per year. Areas in the middle latitudes showed moderate differences, ranging from 0.5 and 1.5 days per year difference. Relatively low differences are visible in Scandinavia and the East

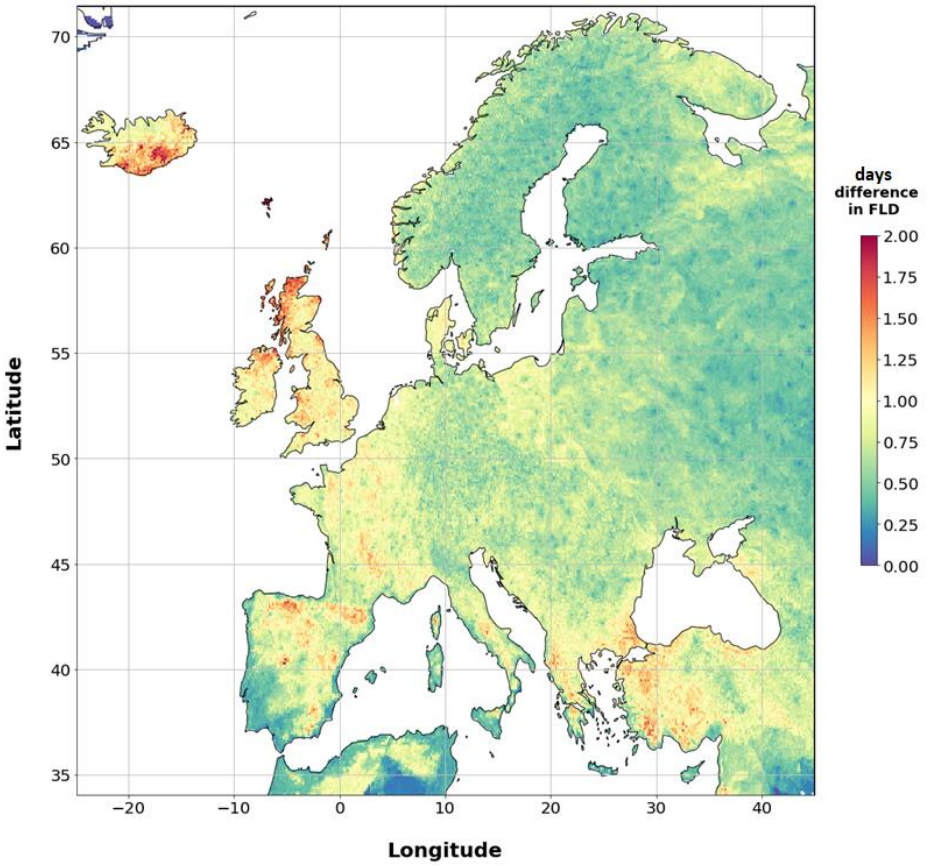


Figure 15. The average difference between the average FLD of the full ensemble and the FLD resulting from the ensemble mean. The blue color corresponds with small differences between the methods, whereas the red color corresponds with large differences.

European Plains in Western Russia, with difference values ranging from 0.25 and 1 days per year difference. The most difference in FLD outputs from the two methods are found in the north of the United Kingdom and in Iceland, as well as in the Spanish Meseta and regions in Greece and Turkey. Differences in these regions could be as high as 2 days difference on average. Areas with high weather station density, such as in Germany, show relatively little difference between the two methods. ANCOVA showed a slight difference in the slope between the ensemble mean (slope of -0.18) and the full ensemble (slope of -0.20) FLD's. However, the difference in slope was non-significant (P = 0.42).

4.3 False spring

Like the first leaf date values, the DI values were calculated for all the cells in the spatial extent of the dataset and all the years in the temporal scope. Figure 16 shows the averaged DI output for the years 2007 (figure 16A) and 2008 (figure 16B). The negative values that are shown with the green color correspond with areas where the FLD occurred after the last frost and the positive values that are shown with the red color correspond with areas where the FLD occurred before the last freeze. As with the FLD values, the variation of DI output is reasonably broad, with the highest damage index values of around 80 and the lowest DI under -40. As shown in figure 16, the DI values vary significantly for different years, showing the erratic nature of false springs. For instance, the DI was positive for most grid cells in Iceland for the year 2007 (figure 16A), whereas those grid cells were negative for the year 2008 (figure 16B). However, there are also recurring spatial patterns in the DI values over the years. Figure 17 shows the average DI values over all the years in the study period. Mountainous regions, such as the Alps, the Pyrenees, and the Caucasus Mountains, on average have very high DI values,

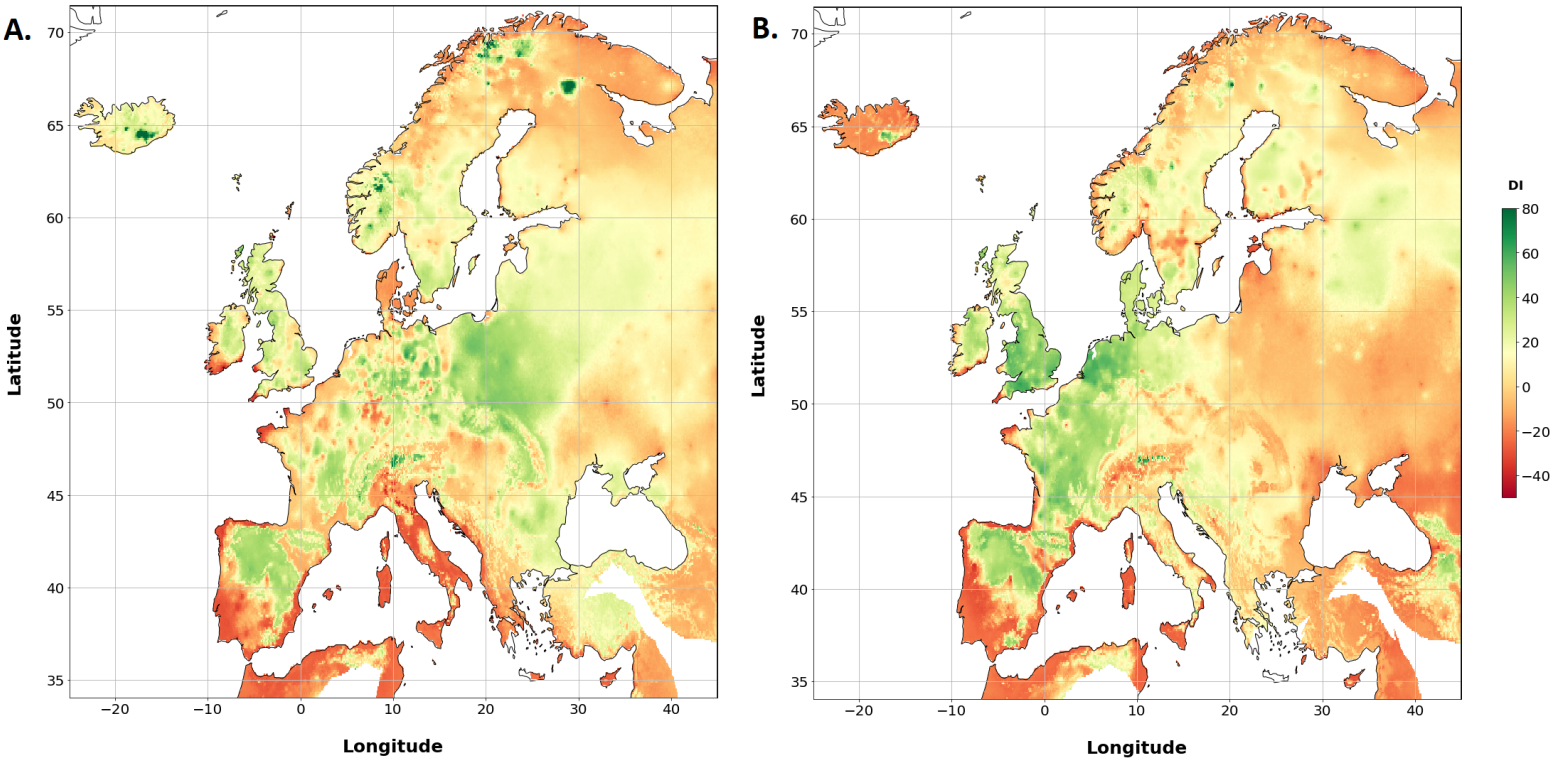


Figure 16. The average DI for the year 2007 (A) and 2008 (B). The negative values that are shown with the green color correspond with areas where the FLD occurred after the last frost and the positive values that are shown with the red color correspond with areas where the FLD occurred before the last freeze. The DI values vary significantly for 2007 and 2008, showing the erratic nature of false springs.

which means that false springs are more likely to occur in these regions. In Iceland, you can clearly see the ice caps emerging in red. The large red dot indicates Vatnajökull, the largest ice cap in Europe. The green areas can be found in the South of Spain and Italy, the Greece Islands and Northern Africa. These areas do not experience freezing weather often. The spatial and temporal average of the DI was 5.5, indicating that false springs are common.

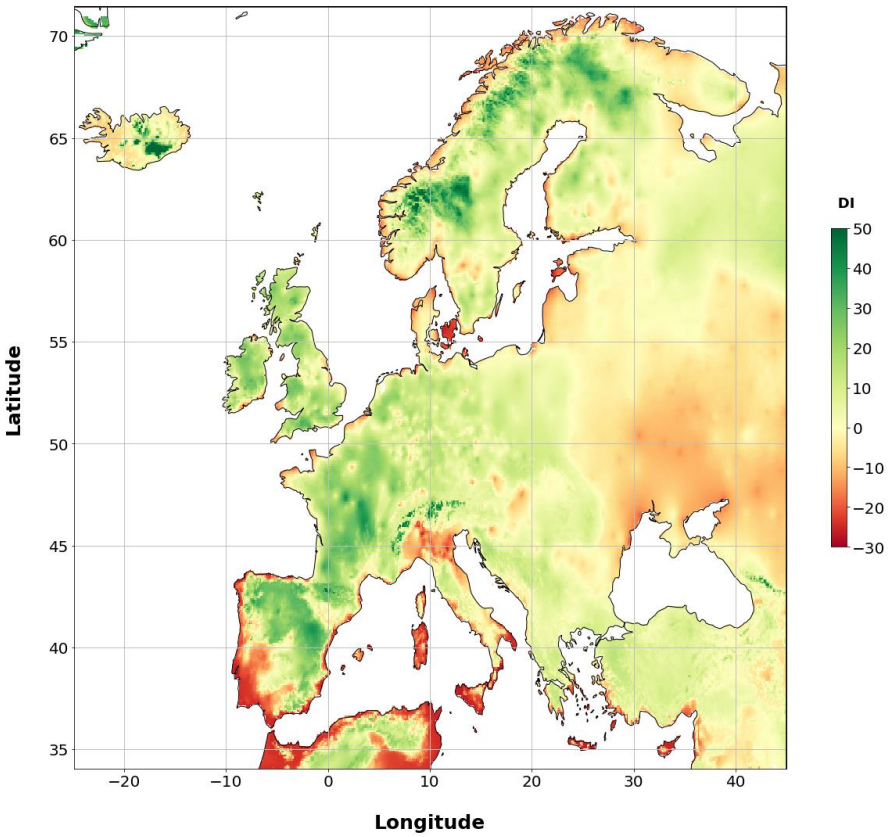


Figure 17. The average DI values over all the years in the study period. The green color depicts areas where false springs are relatively common, whereas in the red areas, false springs are rare.

The variability of the DI values is higher than for the FLD values (see Appendix B). The variability of DI values is especially high in regions where late freezes may occur, such as in Scandinavia and in Alpine regions. These late freezes may occur in some years and not in others, resulting in highly variable DI values. Contrarily, the variability of DI values is low in regions where freezes rarely occur, such as in the south of Spain and parts of Northern Africa. In these regions the variability of DI values is mostly dependent on the FLD values, for which the variability is relatively low in these same regions.

4.3.1 False spring trends

The DI decreases on average with 0.2 days per decade for all grid cells for the full temporal range (see table 3). However, the slope of the trends is positive from 1950 until 1979 (increase of 1.4 days per decade) and negative from 1980 until 2019 (decrease of 0.6 days per decade). Furthermore, the trend is only significant from 1950 until 1979 (p-value of 0.01), and non-significant for the full temporal range and from 1990 until 2019 (p-values of 0.36 and 0.18 respectively). Figure 18 shows the average DI for all grid cells per year and the corresponding trends for the different temporal ranges. As with the FLD, there appears to be a change in slope values from 1980 onwards. The significant positive trend seems to stop around 1980, followed by a more erratic period from 1980 until 2019.

Period	*Slope*	*P-value*	*Std.err.*
1950-1979	0.14	0.01	0.06
1980-2019	-0.06	0.18	0.04
1950-2019	-0.02	0.36	0.02

Table 3. DI slope values for three different temporal ranges, namely 1950-1979, 1980-2019, and 1950-2019. The slope from 1980-2019 is the only significant slope and it is a positive trend.

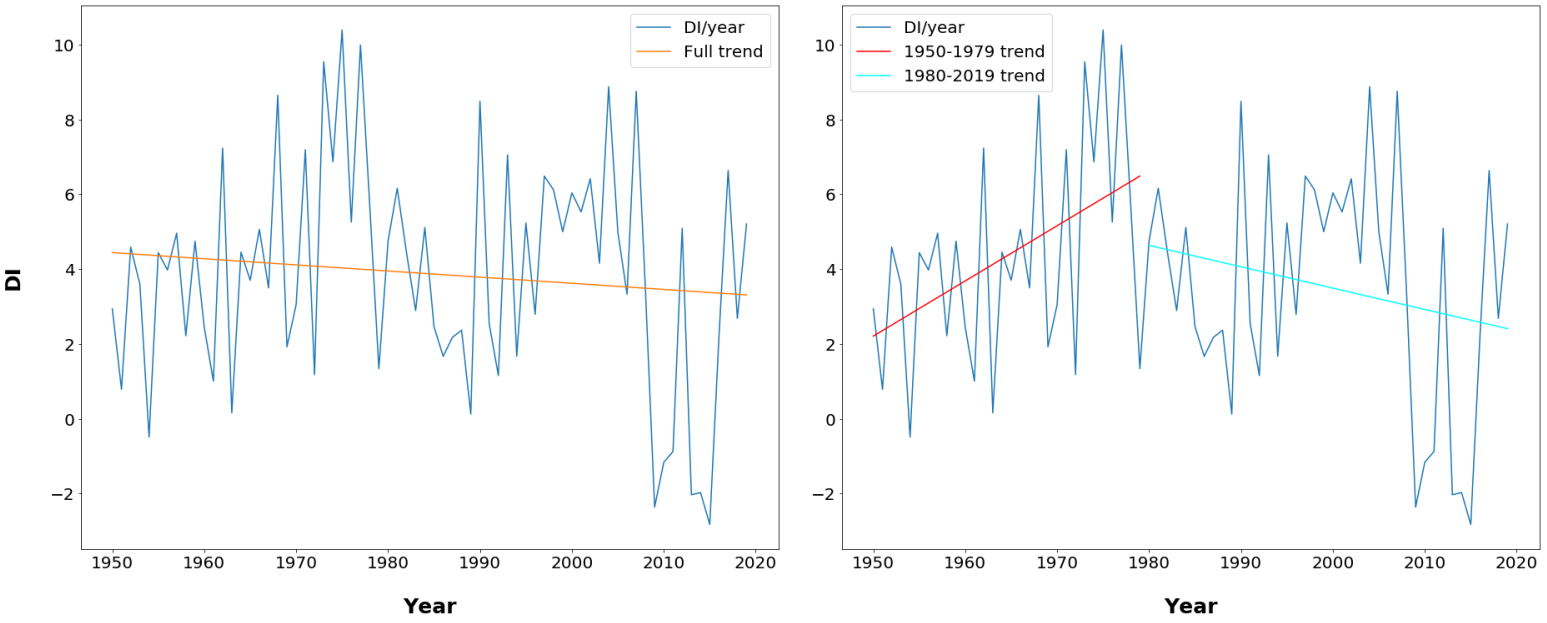


Figure 18. Damage Index (DI) 100 ensemble member average for the year 2007 (A) and the year 2008 (B). The negative values that are shown with the green color correspond with areas where the FLD occurred after the last frost and the positive values that are shown with the red color correspond with areas where the FLD occurred before the last freeze. The spatiotemporal variability of DI is relatively high.

Both the Theil-Sen slopes (figure 19A) and the linear regression slopes (Appendix C) indicate that the DI slope values are spatially highly variable. Positive trends, which means an increase in DI and higher probabilities of false spring occurrences over time, are mainly found the United Kingdom, the Netherlands, Denmark, the Spanish Meseta, and the East European Plains in western Russia. Negative trends, indicating a decrease in DI and lower probabilities of false spring occurrences over time, are mainly found in Iceland, northern Africa, and mountainous regions, such as the Scandinavian Mountains and the Alps. The significance of the trends is high in most regions with clear negative and positive trends, approximating a p-value of 0 (figure 19B). Due to the high spatial variability of DI slope values, there are many

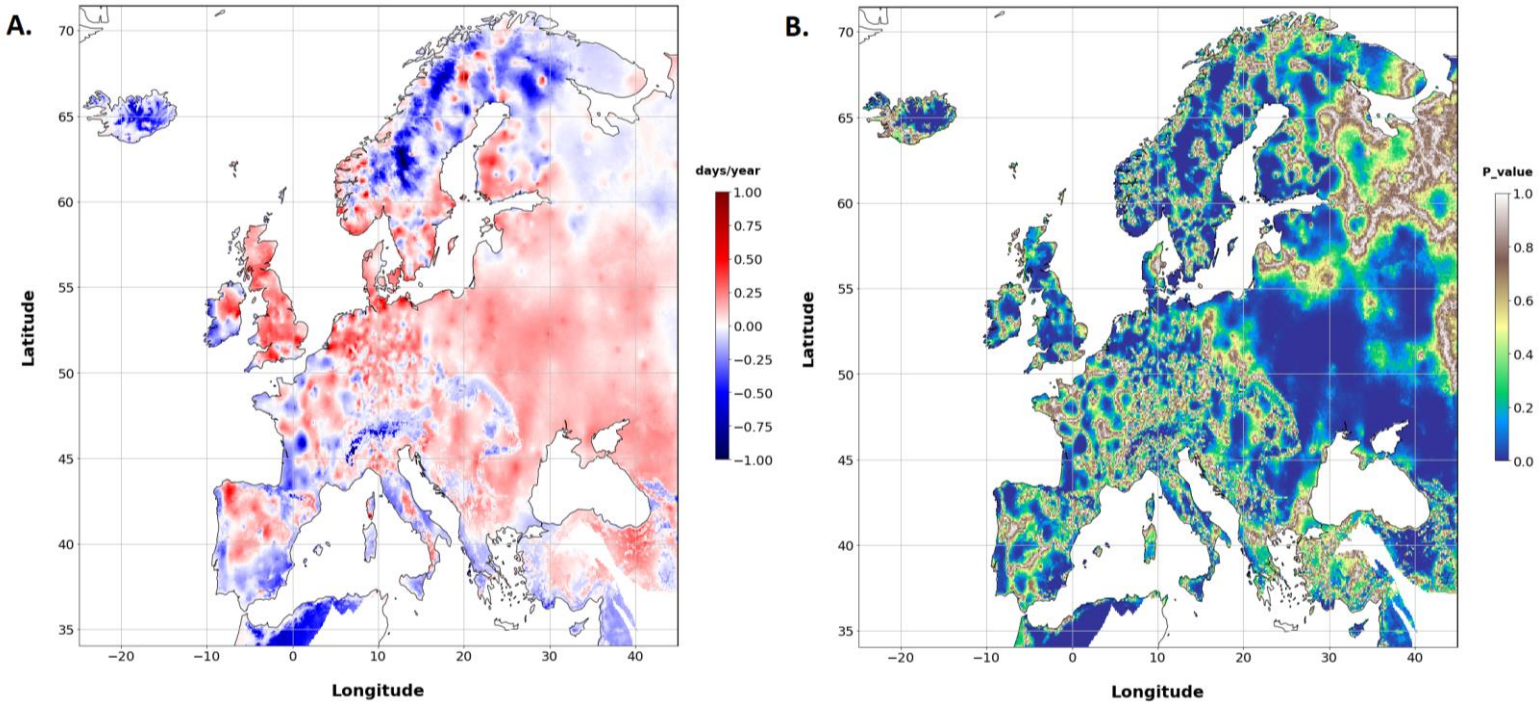


Figure 19. (A) The DI Theil-Sen slope values. The red color indicates areas where there is an increase in DI, whereas the blue color indicates a decrease in DI. The DI slope values are spatially highly variable. (B) The statistical significance of the slope values. Most grid cells with either clear negative or positive slope values have p-values approximating 0, slope values of approximately 0 show relatively high p-values and low significance.

regions are between positive slope regions and negative slope regions. These transition zones have trend values of approximately 0 and relatively high p-values (figure 19B). Appendix D shows the spatiotemporal patterning of slope and significance values for 1950 until 1979 and from 1980 until 2019. From 1950 until 1980, the slope values are mostly positive, especially in western Europe, whereas the slope values are less prominent and more spatially diverse for 1980 until 2019. Interestingly, for many regions the slope values are reversed for the different temporal ranges.

4.3.2 Impact of full ensemble on DI trends

To assess the impact of employing the full temperature ensembles on the DI outputs, the difference between the average DI of the full ensemble and the DI resulting from the ensemble mean was calculated for each year and averaged over the study period (figure 20A). The differences in DI values between the two methods are more prominent than the differences between FLD. As with the FLD differences between the two methods, regions with very early FLD values showed the least difference in DI values between the two methods, approximating an average difference of 0 to 2.5 DI value per year. Most areas in the mainland Europe showed moderate differences, ranging from 5 to 10 DI values difference. Relatively high differences are visible in parts of Scandinavia, Spain, North Africa, the United Kingdom, Iceland, and France, with DI difference values ranging from 12.5 to 20 on average per year. The dots with lower difference values compared to their immediate surroundings, for instance in Scandinavia and Russia, are the locations of weather stations. In contrast with the FLD slopes, the ANCOVA showed a significant change in the DI slope between the ensemble mean and the full ensemble ($P < 0.001$). The slope for the ensemble mean is -0.02 days per year (or a 0.2 DI per decade decrease) and the slope for the full ensemble is 0.05 (or a 0.5 DI per decade increase). Locally, the differences in DI trends for the full ensemble and the ensemble mean calculated with ANCOVA were mostly significant (figure 20B). The regions where station density is relatively high (see figure 22) showed non-significant differences in trends. These mostly include regions in Germany, Scandinavia, and the Caucasus mountains.

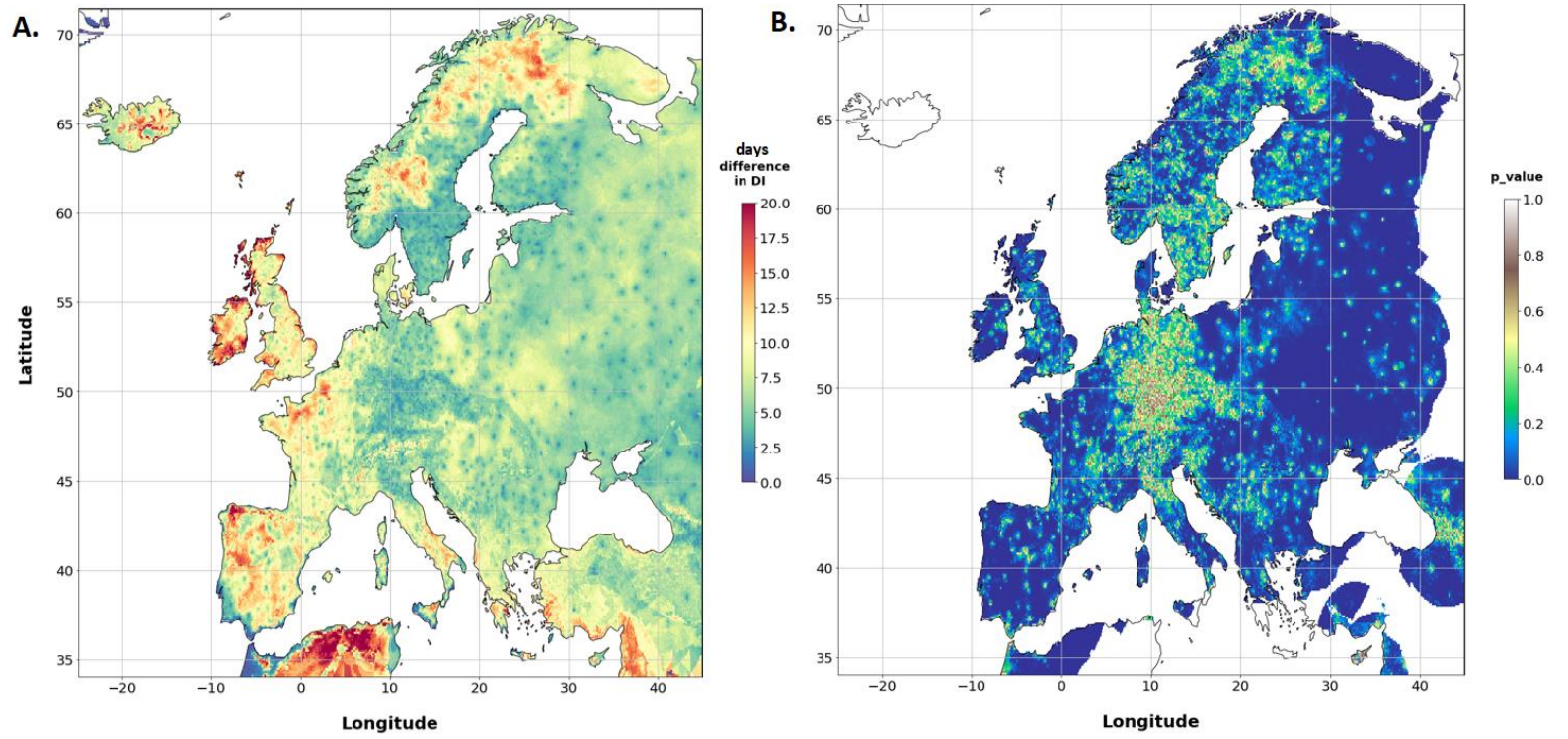


Figure 20. (A) The average absolute difference between the average DI of the full ensemble and the DI resulting from the ensemble mean. The blue color corresponds with small differences, whereas the red color corresponds with big differences. (B) The significance of the difference in DI trends for the full ensemble and the ensemble mean calculated with ANCOVA. The dark blue color corresponds with low and significant p-values, which can be found in most grid cells.

4.4 Uncertainty assessments

4.4.1 FLD uncertainty

The uncertainty in FLD output is expressed as the coefficient of variance (CV) and the 90 percent confidence range (90% CI) over the ensemble members, rendering FLD uncertainty measures for each grid cell for each year. The average CV and 90% CI values over the study period are summarized in figure 21. The average CV (figure 21A) and the average 90% CI values (figure 21B) show similar patterning, with high uncertainty values in the United Kingdom, the Spanish Meseta, Greece, and Turkey, and low uncertainty values in Scandinavia and the East European Plains. However, the average CV is higher at lower latitudes since the CV normalizes variance with the mean and lower latitudes have significantly lower mean FLD values due to warmer climates.

For the uncertainty propagation approximation, the average spread of mean temperature was calculated. Figure 22 shows the average spread for all days prior to the average FLD value, averaged over all years. The average temperature uncertainty is highest in the regions with lower weather station densities, with average spreads values of 3.5 to 5 °C. These regions include Greece, Turkey, the Middle East, North Africa, the East European Plains, Iceland, and central Spain. The average temperature uncertainty is lowest in regions with high weather station densities, with average spreads values of 1 to 2.5 °C. These regions include Germany, Scandinavia, and the Caucasus mountains. The propagation of uncertainty is defined as the uncertainty of FLD for each degree Celsius uncertainty of the average spread (figure 23). There is a high variability in the propagation of temperature uncertainty into FLD uncertainty. The areas where the propagation of uncertainty is the smallest are in the south of Spain and Portugal, and the coastal areas in North Africa. In these areas, for each degree

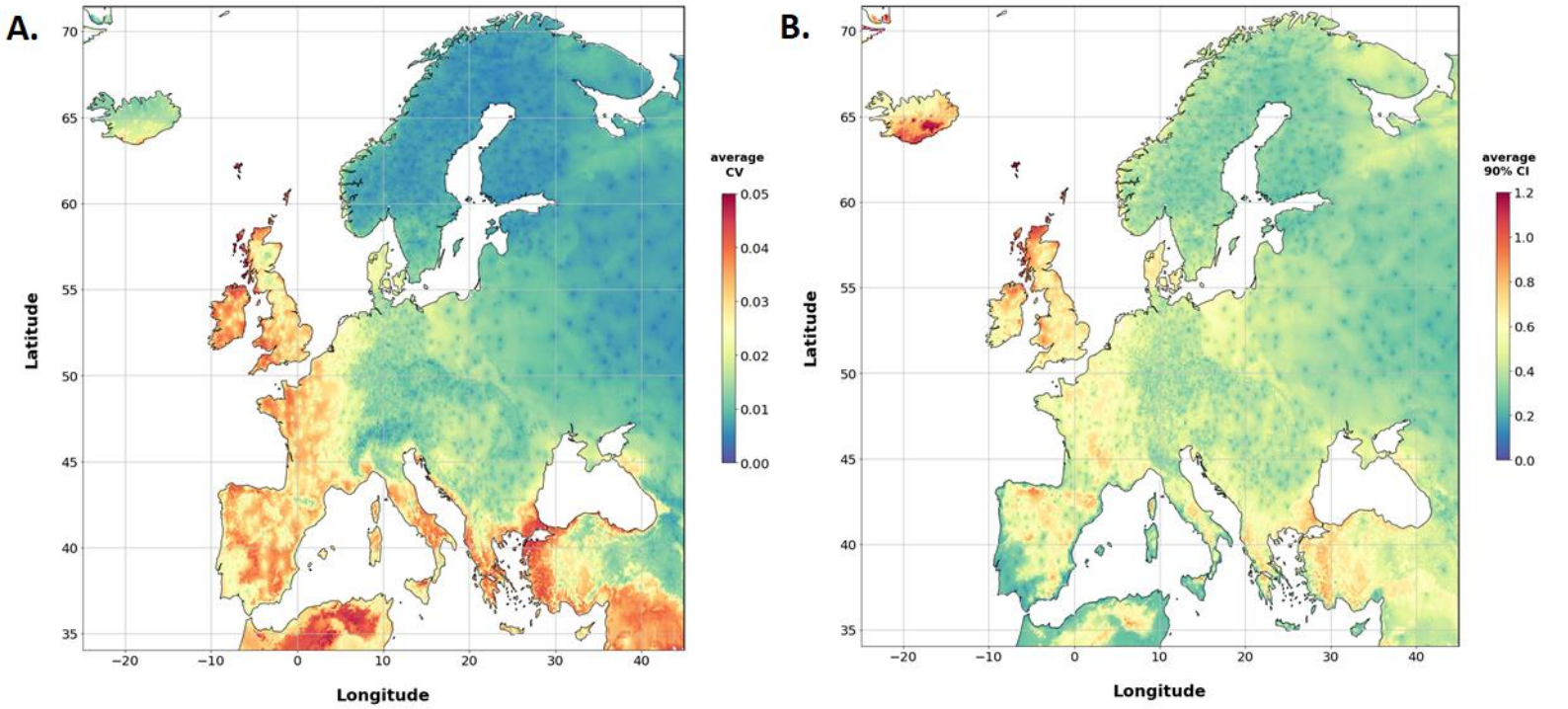


Figure 21. FLD uncertainty approximations. (A) The average coefficient of variance (CV) of FLD over all study years. (B) The average 90 percent confidence interval (90% CI) over all study years. Although the patterns are approximately the same, the average is higher at lower latitudes since the CV normalizes variance with the mean and lower latitudes have significantly lower mean FLD values due to warmer climates.

Celsius average temperature uncertainty, the FLD uncertainty is approximately 0.05 to 0.10 days. The propagation of uncertainty is slightly higher in the East European Plains and areas in the Middle East, where each degree Celsius uncertainty results in an FLD of approximately 0.10 to 0.15 days. The propagation of uncertainty is moderately high in Scandinavia, East Europe (excluding Russia), Turkey, and inland North Africa. In these areas, for each degree Celsius average temperature uncertainty, the FLD uncertainty is approximately 0.15 to 0.20 days. Lastly, the areas with the highest propagation of uncertainty are in western Europe, peaking in the United Kingdom, Iceland, and the Netherlands. In these areas, the FLD uncertainty is roughly 0.2 to 0.4 days for each degree Celsius uncertainty.

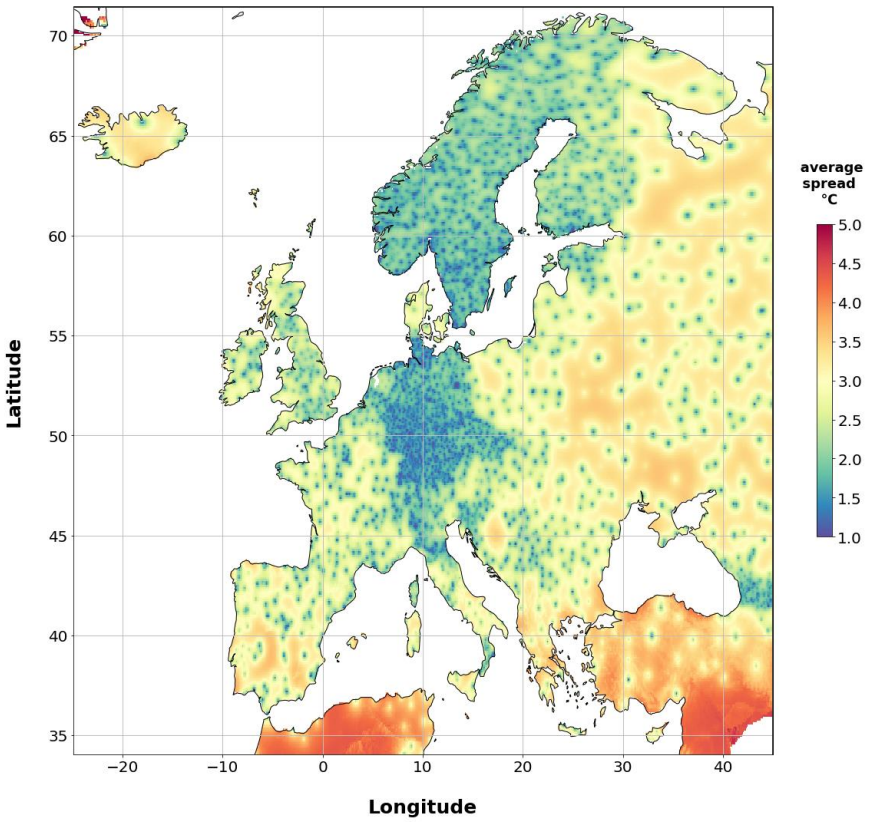


Figure 22. The average spread for all days prior to the average FLD value, averaged over all years. The average temperature uncertainty is highest in the regions with lower weather station densities. The dots that are visible are lower spread values that result from the proximity of a weather station.

4.4.2 False spring uncertainty

The uncertainty in DI values is indicated with the 90% CI. Figure 24 shows the average 90% CI of DI averaged over the study period. The range of the uncertainty interval is higher as compared to the FLD 90% CI with values ranging from 0 to 7 days difference (as compared to 0 to 1.2 days difference). The spatial pattern of DI uncertainty approximates that of the last freeze date uncertainty (Appendix E) because the LFD uncertainty is much higher than the LFD uncertainty, and therefore more influential in the determination of DI uncertainty. The difference between the LFD of ensemble members could be months. The DI uncertainty is lowest in areas where freezes do not or rarely occur, such as the coastal areas of North Africa, the southern coasts of Sicily and Spain. In these areas the uncertainty range is approximately 0 to 1. The East Europe has

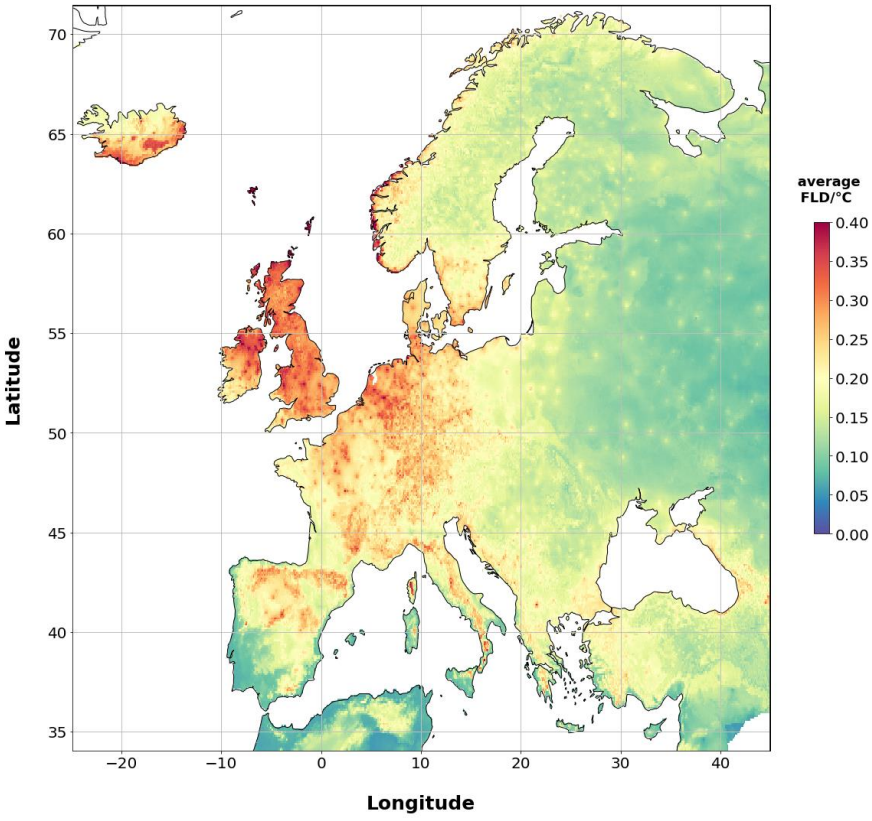


Figure 23. The propagation of temperature uncertainty to FLD uncertainty. The propagation of uncertainty is defined as the uncertainty of FLD for each degree Celsius uncertainty of the average spread.

intermediate uncertainty ranges, with values ranging 2 to 4. The highest uncertainties are found in West Europe, where values can be as high as 7. Mountainous regions, such as the Alps and the Caucasus, have higher uncertainty than their immediate surroundings.

Besides the DI, the binary false spring probability values for a 0-, 7-, 10- and 14-day lag were calculated. Figure 25 shows the average probability values for the different binary false springs. The patterning for the different binary false springs is approximately the same, however, the probability of false spring diminishes with increasing lag times. In general, false springs are likely in West Europe, and specifically in France, the Spanish Meseta, the Scandinavian mountains, and the United Kingdom. The probability of a false spring is low in the East European Plains, and in the areas where freezes rarely occur. Furthermore, coastal areas are less likely to have false springs as compared to inland areas.

The uncertainty of binary false spring is calculated with Shannon's binary entropy function. This function penalizes intermediate p-values where uncertainties are high. The binary false spring uncertainty is low in areas where the false spring probability values are either very low or very high. For instance, the East European Plains have very low probabilities of false springs, which translates to low uncertainties, whereas the Spanish Meseta have high probabilities of false springs and thus also low uncertainties (Appendix F). The uncertainty of binary false springs in general decrease with increasing lag times for areas where the probability of early false springs is already low, for instance in the East European Plains and most coastal areas. Contrarily, the uncertainty of binary false springs generally increases with increasing lag times for areas where the probability of early false springs is high, such as in France and the Spanish Meseta. In general, the uncertainties of earlier binary false springs (0- and 7-day lag) are higher than the uncertainties of later binary false springs (10- and 14-day lag).

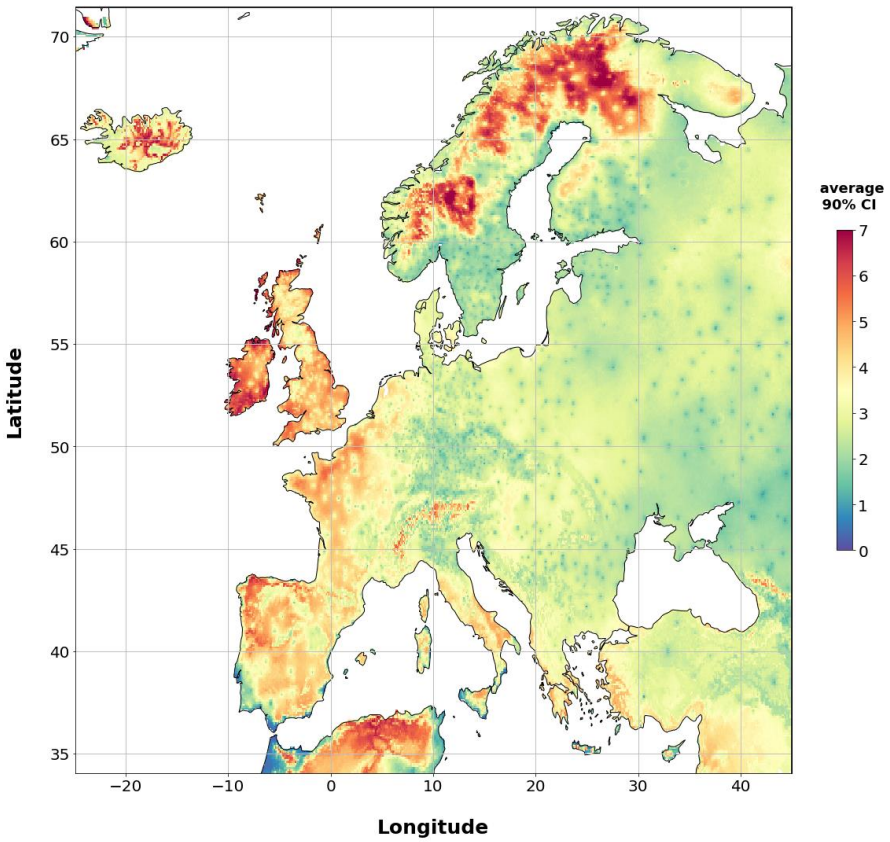


Figure 24. The average 90% CI of DI averaged over the study period. The areas with the red colors have relatively high uncertainty in DI. The highest uncertainties are found in West Europe, where values can be as high as 7. Mountainous regions, such as the Alps and the Caucasus, have higher uncertainty than their immediate surrounding.

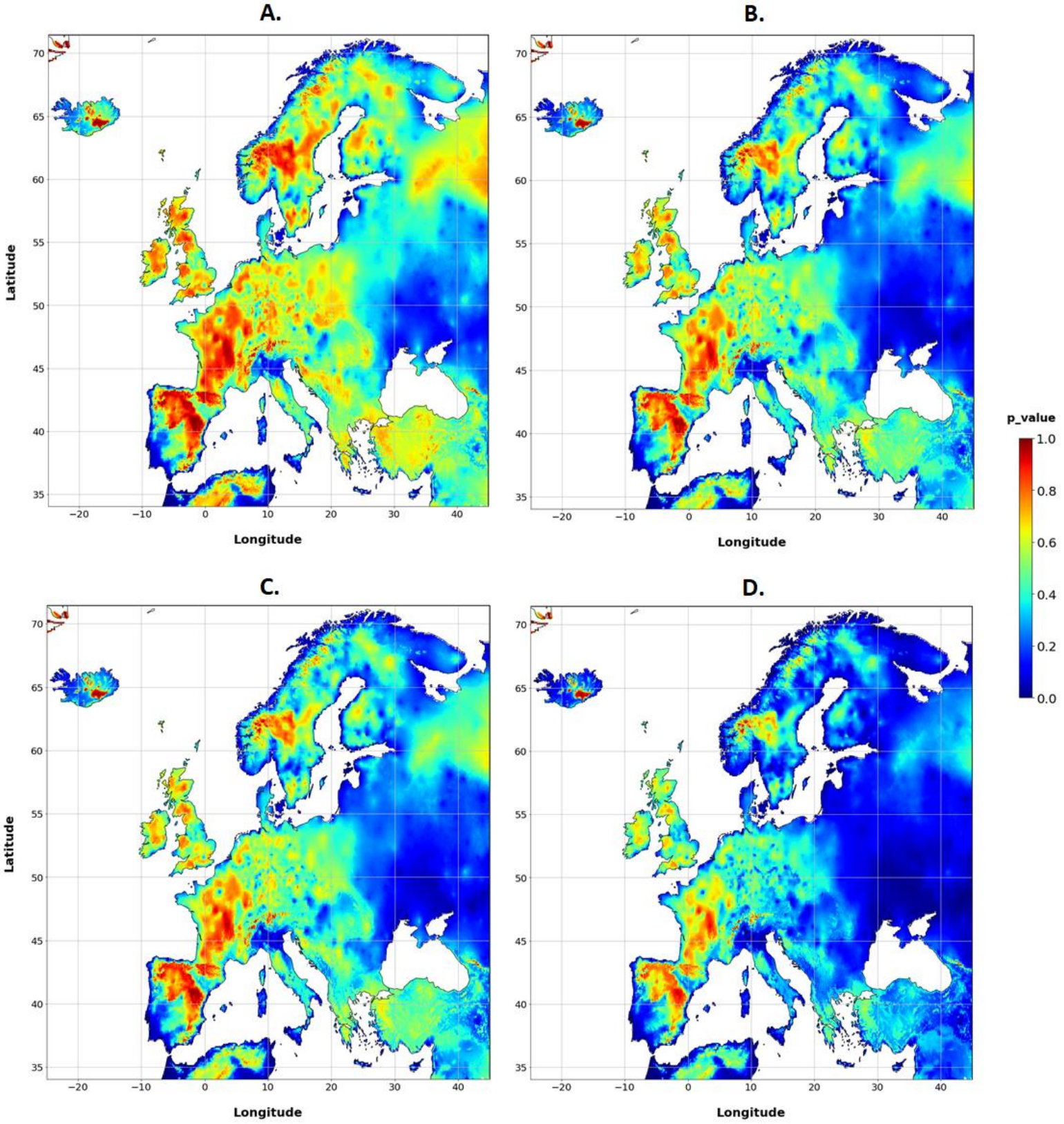


Figure 25. The binary false spring probability values for the (A) 0-, (B) 7-, (C) 10-, and (D) 14-day lag were calculated. The blue color corresponds with low p-values and the red color with high p-values. The patterning for the different binary false springs is approximately the same, however, the probability of false spring diminishes with increasing lag times.

5. Discussion and conclusions

In this fifth chapter, the temperature data and the phenological model that are used in this research will be discussed, and limitations regarding to the data and the model are highlighted. Furthermore, the results from this research will be further elaborated and placed in a broader perspective with relevant scientific literature. From the synthesis, the research question can be answered in the conclusions section.

5.1 Temperature data

The E-OBS temperature data was mostly complete for geographic range. However, some regions did not have sufficient coverage of temperature data and sufficient FLD values for the trend analysis. These regions include areas in central Turkey, the Middle East, and part of North Africa. The spread in average temperature is related to the density of weather stations, as depicted by figure 22. Lower station densities and higher temperature uncertainties were found in the East European Plains, Turkey, and North Africa. The lower uncertainties close to weather stations can clearly be seen in the average temperature spread (figure 22) and several uncertainty approximations (figure 20 until 24) as dots. This nicely illustrates the importance of weather station proximity in uncertainty approximations.

The E-OBS dataset is a parametric ensemble, in which different plausible parameter values are applied to get equally likely realizations based on these different parameter values. These parameter values include the search radius for inclusion of stations that influence the grid cell, impact estimates of homogeneity issues on the quality of input station data, impact of co-variates such as latitude or distance to wat bodies, and the inclusion of station data of which the base period is less precisely known. The 100-member ensemble deals with uncertainty related to these parameters. Errors in the underlying station data can also propagate into the gridded datasets. These errors include incorrect station location information and individual erroneous values or inhomogeneities in the station time series (Cornes et al., 2018; Hofstra et al,, 2010). The E-OBS datasets that are employed in this study do not correct these inhomogeneities. Moreover, the number of stations that are used in the determination of the observational gridded datasets varies over time. These deficiencies may impact the trend analysis based on these temporal gridded datasets.

5.2 Extended spring indices

The SI-x is a widely used and well-parameterized phenological model, which facilitates adequate comparisons across space and time and reduces uncertainty (Ault et al., 2015a; Labe et al., 2017; Zhu et al., 2019). This makes the SI-x model especially useful for scientific analysis of spring onset and false springs. However, the SI-x model has some notable limitations. The current SI-x model uses three shrub species, namely two honeysuckle species and a lilac species, to approximate green-up. Populating the SI-x model with a richer ecology of indicator species could make the SI-x model more comprehensive (Ault et al., 2015b). Alternatively, species-specific models could be designed to ensure more accurate predictions to better understand potential effects on natural and agricultural systems (Allstadt et al., 2015; Mehdipoor et al., 2017). Another possible drawback of the SI-x model is that it reflects temperature variation only. Parameters such as preseason temperature, solar radiation, precipitation, winter chilling, biotic factors, and large scale circulation anomalies are not taken

into account into the predictions (Ge et al., 2014; Ma et al., 2012; Mehdipoor et al., 2016). For instance, in arid regions water is often a limiting factor and precipitation is often correlated with spring onset (Piao et al., 2006). Likewise, warmer winter temperatures in temperate regions could counterintuitively result in delayed spring onset since chilling requirements are not met (Guo et al., 2015), which may result in an overestimation of spring onset advancement in these regions (Zhu et al., 2019). Lastly, the SI-x assumes vegetation homogeneity, whereas in reality, there are geographic variations in vegetation. Moreover, species can adjust phenological responses based on climatic change, resulting in temporally differing responses to similar climatological circumstances.

5.3 Computational solution

The Dask implementation of the extended spring index proved to be a valuable alternative to the conventional MATLAB implementation. The output from both implementations is identical. Moreover, the computation in the Dask environment resulted in less than 38% of computation time than the MATLAB implementation. When running the model for high resolutions, this reduction could result in significantly shorter computation times, enabling researchers to receive their spring onset and false spring output data faster. The Python environment is an open-source platform, enabling easy access and use. MATLAB, on the other hand, is proprietary, closed-source software with expensive licenses. This greatly reduces accessibility to the MATLAB environment for most. Lastly, the Python environment allows for scalability through packages like Dask, whereas MATLAB does not provide this scalability.

Even though the implementation used in this study greatly reduced computation times, the employment of the full ensemble and the full temporal range that was available resulted in a staggering 330 hours of computation. Therefore, further optimizations are desirable when running the model on this volume of data. For instance, the model could be rewritten to ignore Not a Number (NaN) values in the computation by masking the areas that do not have valid temperature values for all ensemble members and all years. For the geographic range in this study, this could theoretically result in over 50% reduction in computation time since over half of the grid cells represented NaN values. Furthermore, even though optimized version of the SI-x model ensures that daylengths are calculated only once for each ensemble member, the model could be rewritten to ensure identical daylengths are calculated once for all ensemble members and years. Lastly, the Dask implementation could be optimized. In the current implementation, some workers finish their computation of the ensemble members faster than others, rendering them idle until the last worker finished its share of the computation. The lag between the completion of computation between the workers could be over a minute, greatly reducing the efficiency of the Dask implementation. To overcome this problem, different ensemble members could be loaded dynamically to the idle workers, reducing overall computation time. Even though these suggested optimizations would further reduce computation times, it is likely that employing high performance computers is desirable when running the SI-x model on full ensembles at high resolution.

5.4 Spring onset predictions

As expected, the average FLD was later in higher latitudes and mountainous areas, where average temperatures are generally lower (figure 11). The across year variability of FLD was spatially highly heterogeneous (figure 12). Especially mid-latitudes in West Europe

showed high year-to-year variability in FLD. The low variability in warmer climates are logical since consistently high temperatures in the beginning of the year result in similarly early FLDs, and thus low variabilities. The relatively low variability in Scandinavia and the East European Plains is somewhat harder to explain. It seems that later spring onsets correspond with lower across-year variabilities. These regions, in general, have shorter daylengths and lower temperatures, resulting in less accumulation of temperature. In turn, this may result mostly short-term growing degree hours and relatively low cumulative count of high synoptic events. The absence or presence of cumulative counts of high synoptic events can impact the FLD significantly (Zhu et al., 2019). The low impact of the cumulative count of high synoptic events may thus explain the lower variabilities in FLD. The high across-year variability in the mid-latitudes in West Europe may be caused by yearly early-year fluctuations of temperatures over and under the base temperature and variable across-year cumulative count of high synoptic events.

In accordance with previous studies (Schwartz et al., 2006; Wu et al., 2016), a significant advancement of spring onset was seen for most regions in Europe (figure 13 and figure 14). West Europe and mountainous regions especially showed sharp decreases in FLD. Temporally, the advancement was most prominent from approximately 1980 onwards as a result of an increase in global temperatures. The trend slopes modelled with the 100-member ensemble did not significantly differ from the trend slopes derived with the ensemble mean. However, the average difference between the two methods is not negligible in most areas, peaking in Iceland and the United Kingdom with an average difference of two days difference between the two methods (figure 15). Moreover, the uncertainty can be quantified by employing the full ensemble. This indicates that even though the trend slopes are not significantly impacted, the influence of working with the 100-member ensemble on FLDs should not be completely disregarded.

5.5 False spring predictions

In areas where freezes rarely occur, false springs were relatively uncommon and average DI values were negative (figure 17). In most regions, and Scandinavia and mountainous areas in particular, the average DI was positive, indicating that freezes often occur after FLD. The geographically dominant increase in DI as a consequence of increasing mean temperature may seem counterintuitive but has been already been reported in historical studies and may be the result of a faster advancement of spring onset as compared to LFDs (Allstadt et al., 2015; Augspurger, 2013; Ma et al., 2019). The variability in DI was higher in areas where (late) freezes are relatively common. Changes in false spring risk depend on both the change in timing of spring onset and last freeze date. In accordance with literature (Allstadt et al., 2015; Augspurger, 2013; Ma et al., 2019; Zhu et al., 2019), the temporal change in false spring risk is spatially highly heterogeneous, with some areas showing increased risk of false spring and other areas decreased risks (figure 19). Interestingly, the DI trends and spatiotemporal patterning from 1950 until 1979 and from 1980 until 2019 are predominantly different (Appendix D). Many areas even show reverse trends for the different temporal windows. In general, from 1950 until 1979 there was a significant increase in false spring risk, whereas, although non-significant, there was a decrease in false spring risk from 1980 until 2019. The trend slopes modelled with the 100-member and the trend slopes derived with the ensemble mean were significantly different. This significant difference was also found for most regions in the geographical extend, especially for grid cells further from weather stations (figure 20B). Moreover, the average difference between the two methods was substantial in most

areas (figure 20A). This indicates that working with the 100-member ensemble has a high impact on both the trends and the average DI.

5.6 Uncertainty assessments

5.6.1 Spring onset uncertainty

Both measures of FLD uncertainty (CV and 90% CI) were relatively low (figure 21). The maximum FLD uncertainty could be found in Iceland and only amounted to an average 90% CI of 1.2 days. The CV was employed to get uncertainty approximations that were normalized for the mean, as it is expected that higher means generally result in higher uncertainties due to longer accumulation periods and uncertainty propagation opportunities. Counterintuitively, both the 90% CI and CV were relatively low for regions where spring onset was late, such as in Scandinavia and the East European Plains. The propagation of uncertainty was calculated as the uncertainty caused by one-degree Celsius average spread (figure 23). Thus, the uncertainty propagation value represents the sensitivity of FLD uncertainty to temperature uncertainty, highlighting regional susceptibility to temperature uncertainty. The uncertainty propagation was highest in West Europe, especially in the mid-latitudes, indicating that these regions are especially susceptible to temperature uncertainty. The FLD uncertainty patterns and the uncertainty propagation patterns resemble the across-year variability of FLD, indicating that it is likely that the same mechanisms are responsible for the uncertainty propagation and the across-year and within-year variability. This could mean that the relative influence of short-term growing degree hour and the cumulative count of high synoptic events in the determination of FLD could be momentous in the uncertainty assessments as well as the across-year variability. Possibly, the high variability and uncertainty in the mid-latitudes of West Europe could be the result of variable cumulative counts of high synoptic events between years and ensemble members. If the temperature accumulation surpasses the synoptic events threshold, the event is counted as a high synoptic event. If regions generally have temperature accumulations that approximate this threshold, the variability between years and ensemble members can be explained since some ensemble members or years will exceed the threshold, whereas others may not. The geographic variability in the relative influence of short-term growing degree hour, the cumulative count of high synoptic events, and daylength (depicted by latitude) in the determination of FLD was high for China (Zhu et al., 2019). Explicitly linking the relative influence of these different factors to uncertainty approximations could be a relevant follow-up study that gives insight in the mechanisms underlying FLD uncertainty.

The methodology used in this thesis to approximate spring onset (and false spring) uncertainty is adequate and produces reliable results. However, the methodology for the uncertainty propagation and the assessment of uncertainty in the trends could be more rigorous and comprehensive. Bayesian methods are often employed for robust uncertainty and uncertainty propagation analysis in environmental sciences (Clark, 2005; Smith et al., 2009). Implementing Bayesian methods could be an improvement in the methodology for future studies quantifying phenological uncertainty from ensemble temperature data, resulting in more robust uncertainty assessments.

5.6.2 False spring uncertainty

The uncertainty of the damage index and binary false springs is substantial, especially in regions where late spring freezes are common, such as in Scandinavia and mountainous areas. In these regions the average 90% CI per year could be as high as 7 days difference in DI (figure 24). Furthermore, the employment of the full ensemble instead of the ensemble mean

had a significant impact on the DI slope. The false spring uncertainty propagated largely from last freeze date uncertainty and not from spring onset uncertainty since the uncertainty of LFD is much higher than the FLD uncertainty and, therefore, more influential in the determination of DI uncertainty (see figure 21 and appendix E). The reason for the substantial uncertainty in last freeze date is explainable from the use of the 100-member ensemble. When some of the ensemble members report that a freeze occurs in late spring and other ensemble members do not show a temperature below the threshold, last freeze dates between different ensemble members could be months apart, resulting in high last freeze date uncertainties. This greatly impacts the false spring assessments and uncertainties associated with these assessments.

While the impact of working with the full ensemble is significant for assessing false spring risk, this uncertainty of last freeze date is essentially overlooked in preceding false spring risk studies with the extended spring indices. Ideally, the uncertainty associated with the last freeze date should be considered and reported in false spring risk research with the employment of the full temperature ensembles. However, since the computational effort of running the full ensemble of temperature data on high-resolution long-term data is significant, this might not be ideal for all prospective studies. Though, computing resources are improving over time and computational challenges related to handling these volumes of data may soon be of the past. Alternatively, the false springs could be reported differently. Perhaps the notion of a false spring with a single temperature threshold is too simplistic and highly susceptible to uncertainty. The false springs could be categorized in different strengths as a range close to the threshold value. Then, if a temperature of -2.1 °C arises, it will not be completely disregarded but will be administered as a slightly less likely/strong last freeze date. The range in which the strength of false springs is calculated could represent the freezing tolerance range across the species which are meant to be represented by the phenological model that is employed. For instance, researchers can gather data on freezing tolerances across relevant species and across populations of these species and use this to change the false spring risk metrics accordingly with the employment of range of freeze tolerance value that can be used to determine the overall strength of the false spring. This approach, however, would have certain limitations. If a stronger frost is followed by a weaker frost the next day, there will be some ambiguity whether the weaker later frost should be addressed or the stronger earlier frost. False spring researchers could address these limitations and adjust metrics that both represent variation in freezing tolerance and remove the high uncertainty related to the binary nature of a single threshold last freeze date. Using a gradual approach instead of a binary, single threshold last freeze date approach could eradicate some of the uncertainty associated with false spring assessments. However, the full ensemble should still be employed to fully capture the uncertainty related to false spring risk assessments.

Defining binary false spring uncertainty with the Shannon binary entropy function was a novelty of this study. The resulting uncertainties of the different lag times are shown in appendix F. Overall, the uncertainty of binary false springs with a lag-time of 0 resulted in the highest uncertainty. The binary false springs with lag-times 7, 10, 14 generally had decreasing uncertainties. However, this pattern was reversed for areas with high initial 0-day lag binary false spring probabilities. Besides the reversal of this uncertainty pattern, no astute patterns or mechanisms could be extracted from the binary false spring assessments. The false spring uncertainty assessments with the DI resulted in more insightful results than the binary false spring uncertainty assessments, thus, we recommend the employment of the DI as the false spring definition for prospective studies assessing false spring risk uncertainty.

5.7 Conclusions

RQ1: How to overcome the computational challenge of handling long-term high-resolution geographical data on a continental-scale?

The Dask implementation of the SI-x model effectively parallelized the computation, making the overall computation more time efficient. Implementing the SI-x model in the Dask environment was relatively straightforward, making it a valuable alternative for future studies. However, when employing the full ensemble, it is recommended to in addition utilize high-performance computers to reduce computation times further.

RQ2: How does the performance the distributed model compare to the performance of the legacy implementation?

The output from the Dask implementation is identical to the output of the legacy implementation in MATLAB. Moreover, the computation time of the Dask implementation is over 2.5 times as fast as the MATLAB implementation. Therefore, the Dask implementation that is employed in this study is a valuable alternative for scientists researching spring onset or false spring with the SI-x model.

RQ3: How does the incorporation of temperature uncertainty impact the spring onset trends?

Even though there was a slightly more prominent decrease in spring onset over the study period with the full ensemble, the inclusion of the full ensemble did not significantly impact the spring onset slope. However, the average difference between the two methods was considerable in some areas, indicating that even though the trend slopes are not significantly impacted, the influence of working with the 100-member ensemble on FLDs should not be completely disregarded.

RQ4: How can the propagation of temperature uncertainty into spring onset uncertainty be quantified?

The propagation of uncertainty is estimated with the spread in FLD values and the spread in average temperature. This resulted in geographic variability of uncertainty propagation estimations. The propagations of temperature uncertainty into FLD uncertainty is highest in the mid-latitudes of West Europe.

RQ5: How does the incorporation of temperature uncertainty impact false spring trends?

The false spring trends were significantly impacted by the employment of the full ensemble. Moreover, the average differences between the false spring approximation from the ensemble mean and the full ensemble were substantial. This highlights the importance of regarding temperature uncertainty in false spring risk estimations.

RQ6: How can the uncertainty of false spring predictions be assessed and quantified?

The uncertainty of the DI and binary false springs is approximated differently. The DI incorporates the variability in all FLD and LFD combinations, whereas the binary false spring uncertainties are only relevant for specific relations between FLD and LFD. The DI uncertainty is estimated as the 90% CI and is highest in regions with prevalent late spring freezes. The uncertainty of binary false springs calculated from the percentage of false spring occurrences

among the 100-member ensemble. From these probability values the uncertainty is defined with an entropy function. The DI uncertainty assessments proved to be more insightful than the binary false spring assessments.

RQ7: How do the uncertainty assessments vary for different concepts of binary false spring?

Uncertainty of binary false springs is low if false springs are extremely likely, such as in the Spanish Meseta, or extremely unlikely, such as in the coastal areas of North Africa. As compared to earlier false springs, later false springs estimations become more uncertain in regions with initially high early false spring probabilities, and less uncertain in regions with initially low early spring probabilities. Uncertainty values are in general relatively high due to the high uncertainty in LFDs.

6. References

Abernathey, R., Paul, K., Hamman, J., Rocklin, M., Lepore, C., Tippett, M., Henderson, N., Seager, R., May, R., Vento, D. D. (2017). Pangeo NSF Earthcube Proposal.

Allstadt, A. J., Vavrus, S. J., Heglund, P. J., Pidgeon, A. M., Thogmartin, W. E., & Radeloff, V. C. (2015). Spring plant phenology and false springs in the conterminous US during the 21st century. *Environmental Research Letters*, *10*(10).

Augspurger, C. K. (2013). Reconstructing patterns of temperature, phenology, and frost damage over 124 years: Spring damage risk is increasing. *Ecology*, *94*(1), 41–50.

Ault, T. R., Henebry, G. M., De Beurs, K. M., Schwartz, M. D., Betancourt, J. L., & Moore, D. (2013). The false spring of 2012, earliest in North American record. *Eos*, *Transactions American Geophysical Union 94*(20), 181–182.

Ault, Toby R., Schwartz, M. D., Zurita-Milla, R., Weltzin, J. F., & Betancourt, J. L. (2015a). Trends and natural variability of spring onset in the coterminous united states as evaluated by a new gridded dataset of spring indices. *Journal of Climate*, *28*(21), 8363–8378.

Ault, Toby R., Zurita-Milla, R., & Schwartz, M. D. (2015b). A Matlab© toolbox for calculating spring indices from daily meteorological data. *Computers & Geosciences*, *83*, 46–53.

Belmecheri, S., Babst, F., Hudson, A. R., Betancourt, J., & Trouet, V. (2017). Northern Hemisphere jet stream position indices as diagnostic tools for climate and ecosystem dynamics. *Earth Interactions*, *21*(8), 1–23.

Bock, A., Sparks, T. H., Estrella, N., Jee, N., Casebow, A., Schunk, C., Leuchner, M., Menzel, A. (2014). Changes in first flowering dates and flowering duration of 232 plant species on the island of Guernsey. *Global Change Biology*, *20*(11), 3508–3519.

Cannell, M. G. R., & Smith, R. I. (1983). Thermal Time, Chill Days and Prediction of Budburst in Picea sitchensis. *The Journal of Applied Ecology*, *20*(3), 951-963.

Caradonna, P. J., & Bain, J. A. (2016). Frost sensitivity of leaves and flowers of subalpine plants is related to tissue type and phenology. *Journal of Ecology*, *104*(1), 55–64.

Cayan, D. R., Kammerdiener, S. A., Dettinger, M. D., Caprio, J. M., & Peterson, D. H. (2001). Changes in the Onset of Spring in the Western United States. *Bulletin of the American Meteorological Society*, *82*(3), 399–416.

Chamberlain, C. J., Cook, B. I., García de Cortázar-Atauri, I., & Wolkovich, E. M. (2019). Rethinking false spring risk. *Global Change Biology*, *25*(7), 2209–2220.

Charrier, G., Bonhomme, M., Lacointe, A., & Améglio, T. (2011). Are budburst dates, dormancy and cold acclimation in walnut trees (Juglans regia L.) under mainly genotypic or environmental control? *International Journal of Biometeorology*, *55*(6), 763–774.

Clark, J. S. (2005). Why environmental scientists are becoming Bayesians. *Ecology Letters*, *8*(1), 2–14.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. D. (2018). An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets. *Journal of Geophysical Research: Atmospheres*, *123*(17), 9391–9409.

Crimmins, T. M., Marsh, R. L., Switzer, J. R., Crimmins, M. A., Gerst, K. L., Rosemartin, A. H., & Weltzin, J. F. (2017). USA National Phenology Network Gridded Products Documentation: U.S. Geological Survey Open-File Report 2017–1003, 27p.

Czernecki, B., Nowosad, J., & Jabłońska, K. (2018). Machine learning modeling of plant phenology based on coupling satellite and gridded meteorological dataset. *International Journal of Biometeorology*, *62*(7), 1297–1309.

Dai, J., Wang, H., & Ge, Q. (2014). The spatial pattern of leaf phenology and its response to climate change in China. *International Journal of Biometeorology*, *58*(4), 521–528.

Dai, W., Jin, H., Zhang, Y., Liu, T., & Zhou, Z. (2019). Detecting temporal changes in the temperature sensitivity of spring phenology with global warming: Application of machine learning in phenological model. *Agricultural and Forest Meteorology*, *279*(July), 1-14.

Dask Development Team: Dask: Library for dynamic task scheduling, 2016.

Dietze, M. C., Fox, A., Beck-Johnson, L. M., Betancourt, J. L., Hooten, M. B., Jarnevich, C. S., Keitt, T. H., Kenney, M. A., Laney, C. M., Larsen, L. G., Loeschel, H. W., Lunch, C. K., Pijanowski, B. C., Randerson, J. T., Read, E. K., Tredennick, A. T., Vargas, R., Weathers, K. C., White, E. P. (2018). Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences U.S.A.*, *115*(7), 1424–1432.

Downey, A. B. (2015). *Think Python: How to think like a computer scientist*. Retrieved from http://dl.acm.org/citation.cfm?id=1593015%5Cnpapers2://publication/uuid/32C45786-BC3A-4566-A202-0BA4F4DC5410

Dragoni, D., Schmid, H. P., Wayson, C. A., Potter, H., Grimmond, C. S. B., & Randolph, J. C. (2011). Evidence of increased net ecosystem productivity associated with a longer vegetated season in a deciduous forest in south-central Indiana, USA. *Global Change Biology*, *17*(2), 886–897.

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *BioScience*, *68*(8), 563–576.

Fernandes, R., & Leblanc, S. G. (2005). Parametric (modified least squares) and non-parametric (Theil-Sen) linear regressions for predicting biophysical parameters in the presence of measurement errors. *Remote Sensing of Environment*, *95*(3), 303–316.

Fu, Y. H., Zhao, H., Piao, S., Peaucelle, M., Peng, S., Zhou, G., Ciais, P., Huang, M., Menzel, A., Peñuelas, J., Song, Y., Vitasse, Y., Zeng, Z., Janssens, I. A. (2015). Declining global warming effects on the phenology of spring leaf unfolding. *Nature*, *526*(7571), 104–107.

Ge, Q., Wang, H., & Dai, J. (2014). Simulating changes in the leaf unfolding time of 20 plant species in China over the twenty-first century. *International Journal of Biometeorology*, *58*(4), 473–484.

Gu, L., Hanson, P. J., Post, W. Mac, Kaiser, D. P., Yang, B., Nemani, R., Pallardy, S. G., Meyers, T. (2008). The 2007 Eastern US spring freeze: increased cold damage in a warming world? *BioScience*, *58*(3), 253–262.

Gu, S. (2016). Growing degree hours - a simple, accurate, and precise protocol to approximate growing heat summation for grapevines. *International Journal of Biometeorology*, *60*(8), 1123–1134.

Guo, L., Dai, J., Wang, M., Xu, J., & Luedeling, E. (2015). Responses of spring phenology in temperate zone trees to climate warming: A case study of apricot flowering in China. *Agricultural and Forest Meteorology*, *201*, 1–7.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., & New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *Journal of Geophysical Research Atmospheres*, *113*(D20).

Helsel, D. R., & Hirsch, R. M. (1992). *Statistical Methods in Water Resources*. New York: Elsevier (Vol. 49).

Hofstra, N., New, M., & McSweeney, C. (2010). The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Climate Dynamics*, *35*(5), 841–858.

Hoyer, S., & Hamman, J. J. (2017). xarray: N-D labeled Arrays and Datasets in Python. *Journal of Open Research Software*, *5*.

Izquierdo-Verdiguier, E., Zurita-Milla, R., Ault, T. R., & Schwartz, M. D. (2018). Development and analysis of spring plant phenology products: 36 years of 1-km grids over the conterminous US. *Agricultural and Forest Meteorology*, *262*(June), 34–41.

Jeong, S. J., Ho, C. H., Gim, H. J., & Brown, M. E. (2011). Phenology shifts at start vs. end of growing season in temperate vegetation over the Northern Hemisphere for the period 1982-2008. *Global Change Biology*, *17*(7), 2385–2399.

Karl, T. R., Arguez, A., Huang, B., Lawrimore, J. H., McMahon, J. R., Menne, M. J., Peterson, T. C., Vose, R. S., Zhang, H. M. (2015). Possible artifacts of data biases in the recent global surface warming hiatus. *Science*, *348*(6242), 1469–1472.

Kellermann, J. L., & van Riper, C. (2015). Detecting mismatches of bird migration stopover and tree phenology in response to changing climate. *Oecologia*, *178*(4), 1227–1238.

Labe, Z., Ault, T., & Zurita-Milla, R. (2017). Identifying anomalously early spring onsets in the CESM large ensemble project. *Climate Dynamics*, *48*(11–12), 3949–3966.

LaDeau, S. L., Han, B. A., Rosi-Marshall, E. J., & Weathers, K. C. (2017). The next decade of big data in ecosystem science. *Ecosystems*, *20*(2), 274–283.

Lieth, H. (1974). *Phenology and seasonality modeling*. Springer-Verlag, Berlin.

Ma, Q., Huang, J. G., Hänninen, H., & Berninger, F. (2019). Divergent trends in the risk of spring frost damage to trees in Europe with recent warming. *Global Change Biology*, *25*(1), 351–360.

Ma, T., & Zhou, C. (2012). Climate-associated changes in spring plant phenology in China. *International Journal of Biometeorology*, *56*(2), 269–275.

McCabe, G. J., & Palecki, M. A. (2006). Multidecadal climate variability of global lands and oceans. *International Journal of Climatology*, *26*(7), 849–865.

Mehdipoor, H., Izquierdo-Verdiguier, E., & Zurita-Millla, R. (2017). Continental-scale monitoring and mapping of false spring a cloud computing solution. In *Proceedings of the 14th International Conference on GeoComputation Geospatial Information Sciences 2017, Leeds, United Kingdom* (pp. 5p). University of Leeds.

Mehdipoor, Hamed, Izquierdo-Verdiguier, E., & Zurita-Milla, R. (2016). *Revisiting the extended spring indices using gridded weather data and machine learning* : *abstract.* Abstract from EGU General Assembly 2016, Vienna, Austria.

Mehdipoor, Hamed, Zurita-Milla, R., Augustijn, E. W., & Izquierdo-Verdiguier, E. (2020). Exploring differences in spatial patterns and temporal trends of phenological models at continental scale using gridded temperature time-series. *International Journal of Biometeorology*, *64*(3), 409–421.

Mehdipoor, Hamed, Zurita-Milla, R., Augustijn, E. W., & Van Vliet, A. J. H. (2018). Checking the consistency of volunteered phenological observations while analysing their synchrony. *ISPRS International Journal of Geo-Information*, *7*(12), 487.

Menzel, A., & Fabian, P. (1999). Growing season extended in Europe. *Nature*, *397*(6721), 659.

Menzel, A., Sparks, T. H., Estrella, N., Koch, E., Aaasa, A., Ahas, R., Alm-Kübler, K., Bissolli, P., Braslavská, O., Briede, A., Chmielewski, F. M., Crepinsek, Z., Curnel, Y., Dahl, Å., Defila, C., Donnelly, A., Filella, Y., Jatczak, K., Måge, F., … Zust, A. (2006). European phenological response to climate change matches the warming pattern. *Global Change Biology*, *12*(10), 1969–1976.

Peñuelas, J., & Filella, I. (2001). Responses to a warming world, *294*(October), 793–795.

Peñuelas, J., Filella, I., & Comas, P. (2002). Changed plant and animal life cycles from 1952 to 2000 in the Mediterranean region. *Global Change Biology*, *8*(6), 531–544.

Peterson, A. G., & Abatzoglou, J. T. (2014). Observed changes in false springs over the contiguous United States. *Geophysical Research Letters*, *41*(6), 6413–6419.

Piao, S., Fang, J., Zhou, L., Ciais, P., & Zhu, B. (2006). Variations in satellite-derived phenology in China's temperate vegetation. *Global Change Biology*, *12*(4), 672–685.

Piao, S., Liu, Q., Chen, A., Janssens, I. A., Fu, Y., Dai, J., Liu, L., Lian, X., Shen, M., Zhu, X. (2019). Plant phenology and global climate change: Current progresses and challenges. *Global Change Biology*, *25*(6), 1922–1940.

Sarvas, R. (1974). *Investigations on the annual cycle of development of forest trees: II. Autumn dormancy and winter dormancy*. Communicationes Instituti Forestalis Fenniae.

Schwartz, M. D. (2003). *Phenology: An Integrative Environmental Science*. Dordrecht: Kluwer Academic Publishers.

Schwartz, M. D., Ahas, R., & Aasa, A. (2006). Onset of spring starting earlier across the Northern Hemisphere. *Global Change Biology*, *12*(2), 343–351.

Schwartz, M. D., Ault, T. R., & Betancourt, J. L. (2013). Spring onset variations and trends in the continental United States: past and regional assessment using temperature-based indices. *International Journal of Climatology*, *33*(13), 2917–2922.

Schwartz, M. D., & Marotz, G. A. (1988). Synoptic events and spring phenology. *Physical Geography*, *9*(2), 151–161.

Schwartz, M. D., & Reiter, B. E. (2000). Changes in North American spring. *International Journal of Climatology*, *20*(8), 929–932.

Schweiger, O., Settele, J., Kudrna, O., Klotz, S. & Kühn, I. (2008). Climate change can cause spatial mismatch of trophica. *Ecology*, *89*(12), 3472–3479.

Scully, R. A. (2010). *Intercomparison of PRISM and DAYMET temperature interpolation from 1980 to 2003.* Utah State University, All Graduate Theses and Dissertations, *578*.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*(3), 379-423.

Smith, R. L., Tebaldi, C., Nychka, D., & Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, *104*(485), 97–116.

Stöckli, R., & Vidale, P. L. (2004). European plant phenology and climate as seen in a 20-year AVHRR land-surface parameter dataset. *International Journal of Remote Sensing*, *25*(17), 3303–3330.

Strimbeck, G. R., Schaberg, P. G., Fossdal, C. G., Schröder, W. P., & Kjellsen, T. D. (2015).

Extreme low temperature tolerance in woody plants. *Frontiers in Plant Science*, *6*(October), 884.

Taylor, S. D., & White, E. P. (2020). Automated data-intensive forecasting of plant phenology throughout the United States. *Ecological Applications*, *30*(1), 1–10.

Theil, H. (1992). A rank-invariant method of linear and polynomial regression analysis. In *Henri Theil's contributions to economics and econometrics* (pp. 345-381). Springer, Dordrecht.

Wang, H., Dai, J., Rutishauser, T., Gonsamo, A., Wu, C., & Ge, Q. (2018). Trends and variability in temperature sensitivity of lilac flowering phenology. *Journal of Geophysical Research: Biogeosciences*, *123*(3), 807–817.

Wu, X., Zurita-Milla, R., & Kraak, M. J. (2016). A novel analysis of spring phenological patterns over Europe based on co-clustering. *Journal of Geophysical Research: Biogeosciences*, *121*(6), 1434–1448.

Xu, J., Heue, K.-P., Loyola, D., & Efremenko, D. (2019). The Pangeo big data ecosystem. *2019 Conference on Big Data from Space (BiDS'19)*, 165–168.

Zhou, L., Tucker, C. J., Kaufmann, R. K., Slayback, D., Shabanov, N. V., & Myneni, R. B. (2001). Variations in northern vegetation activity inferred from satellite data of vegetation index during 1981 to 1999. *Journal of Geophysical Research: Atmospheres*, *106*(D17), 20069–20083.

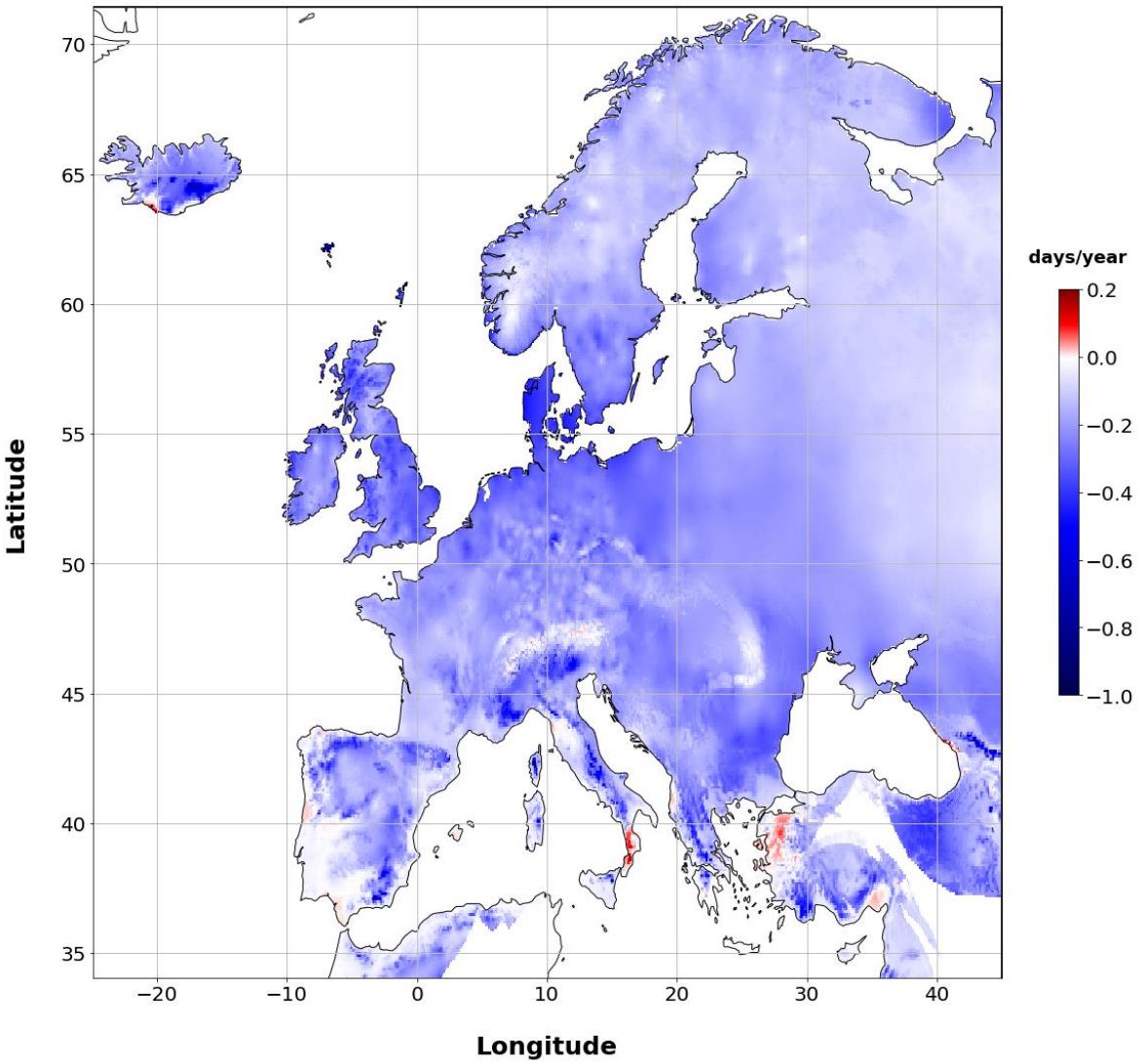
Zhu, L., Meng, J., Li, F., & You, N. (2019). Predicting the patterns of change in spring onset and false springs in China during the twenty-first century. *International Journal of Biometeorology*, *63*(5), 591–606.

Zumwald, M., Knüsel, B., Baumberger, C., Hadorn, G. H., Bresch, D. N., & Knutti, R. (2020). Understanding and assessing uncertainty of observational climate datasets for model evaluation using ensembles. *WIREs Climate Change,* e654.

Zurita-Milla, R., Goncalves, R., Izquierdo-Verdiguier, E., & Ostermann, F. (2019). Exploring spring onset at continental scales: mapping phenoregions and correlating temperature and satellite-based phenometrics. *IEEE Transactions on Big Data*.

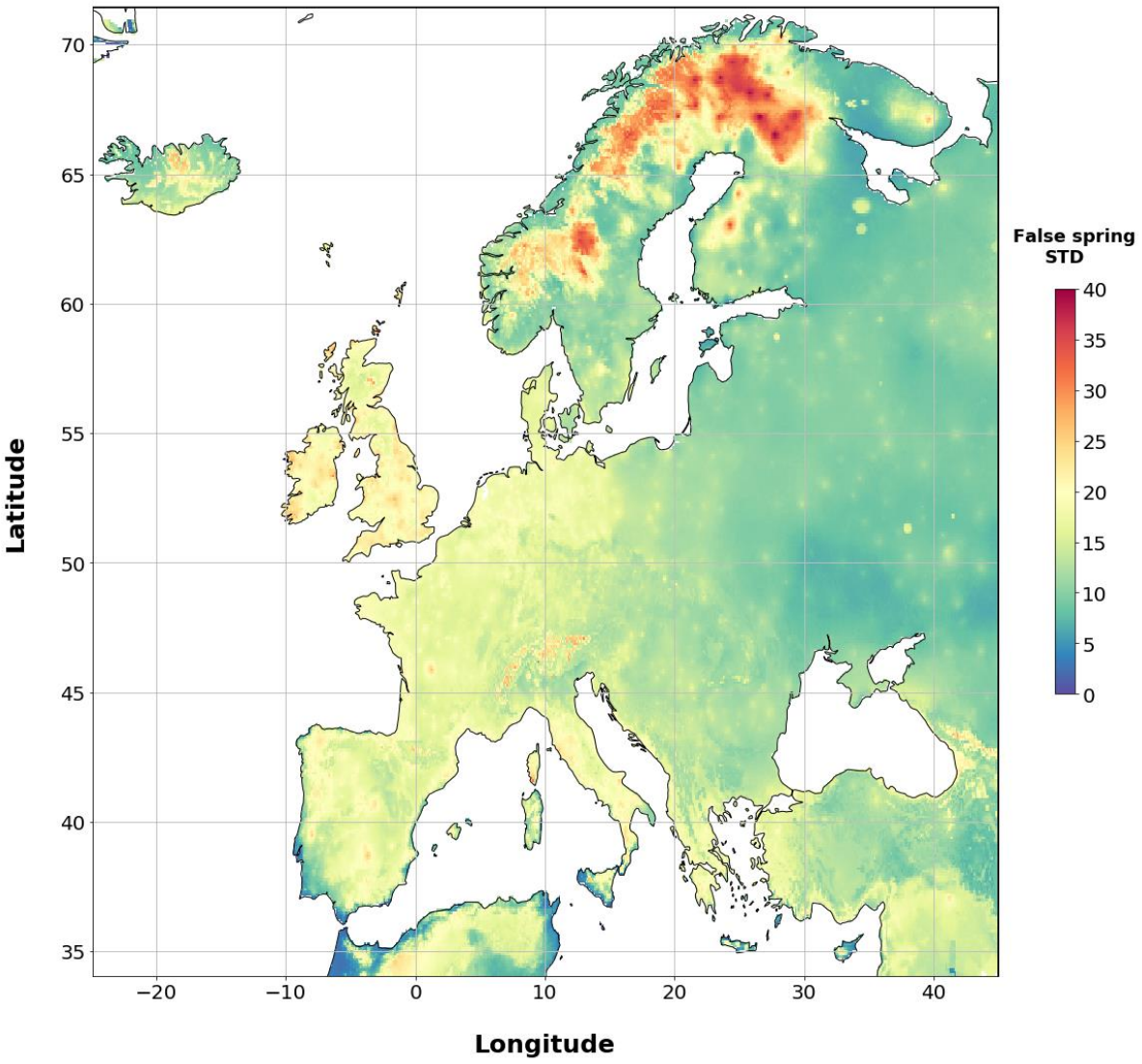
7. Appendix

A. Temporal FLD trends from linear regression



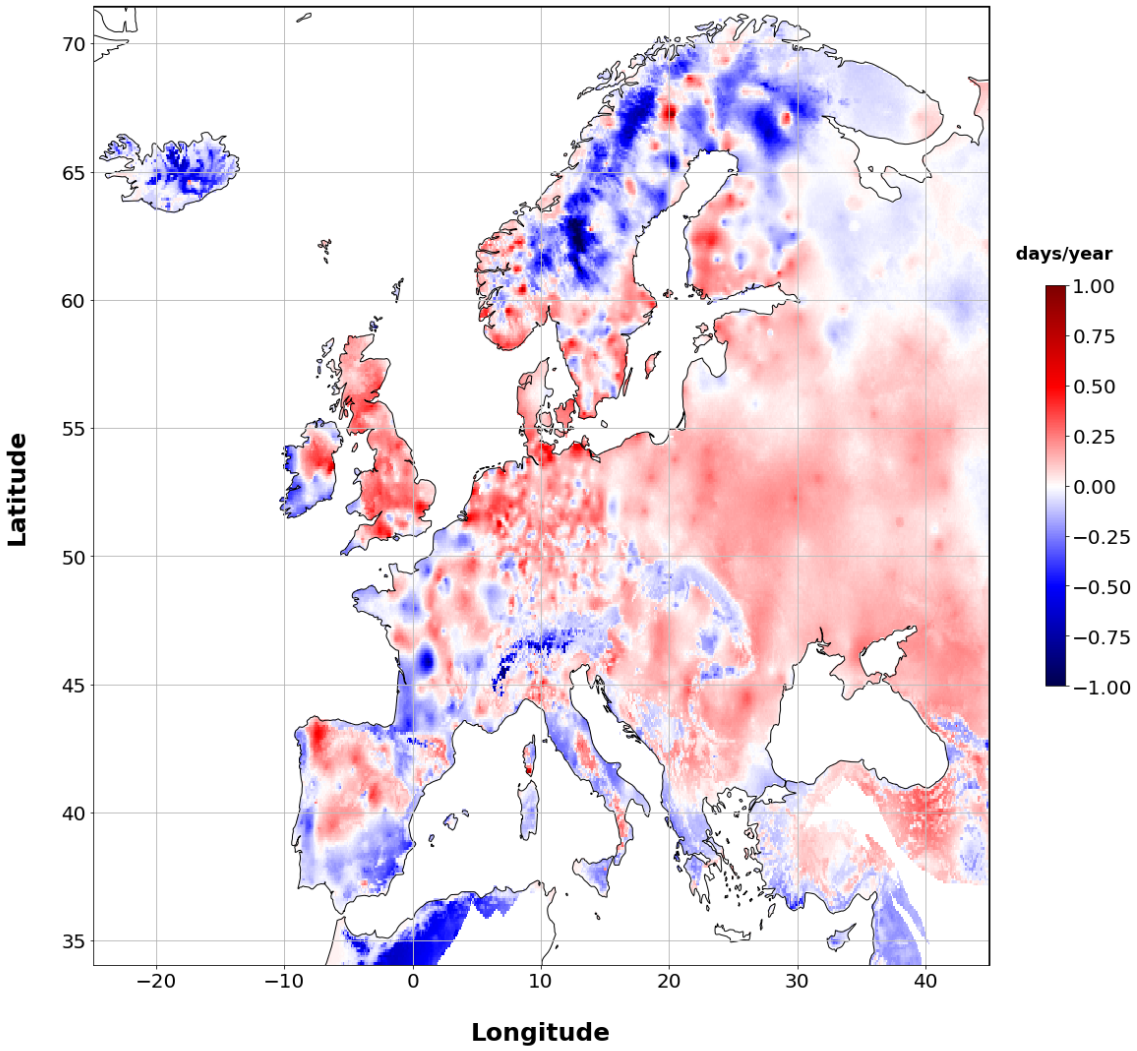
Appendix A. The linear regression FLD slope values for 1950 until 2019. Most grid cells show negative trend values in blue, indicating an advancement of spring onset.

B. The variability of the DI values over the years



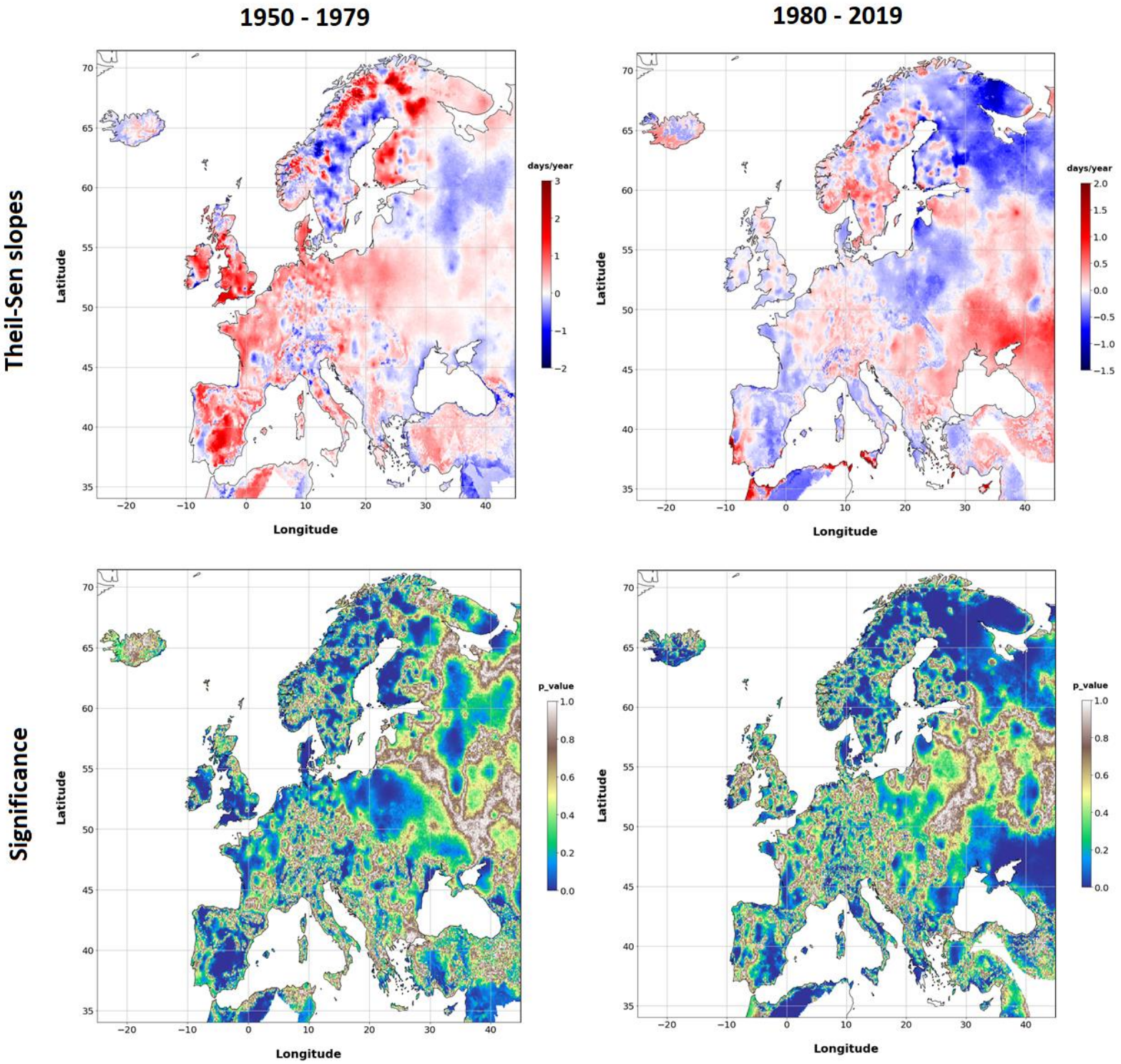
Appendix B. The standard deviation of the DI values for 1950 until 2019. Most grid cells show moderate standard deviation values in yellow and green. Warmer climate regions, such as areas in the south of Spain and areas in northern Africa, show low variability in DI values. These areas show low variability as freezing rarely occurs in these regions and FLD values are more or less constant. Areas in Scandinavia and mountainous regions, such as the Alps and the Caucasus, show high variability. This may be caused by the occurrence of late freezes, which may vary on a year-to-year basis, resulting in a high variability of DI values.

C. Temporal DI trends from linear regression



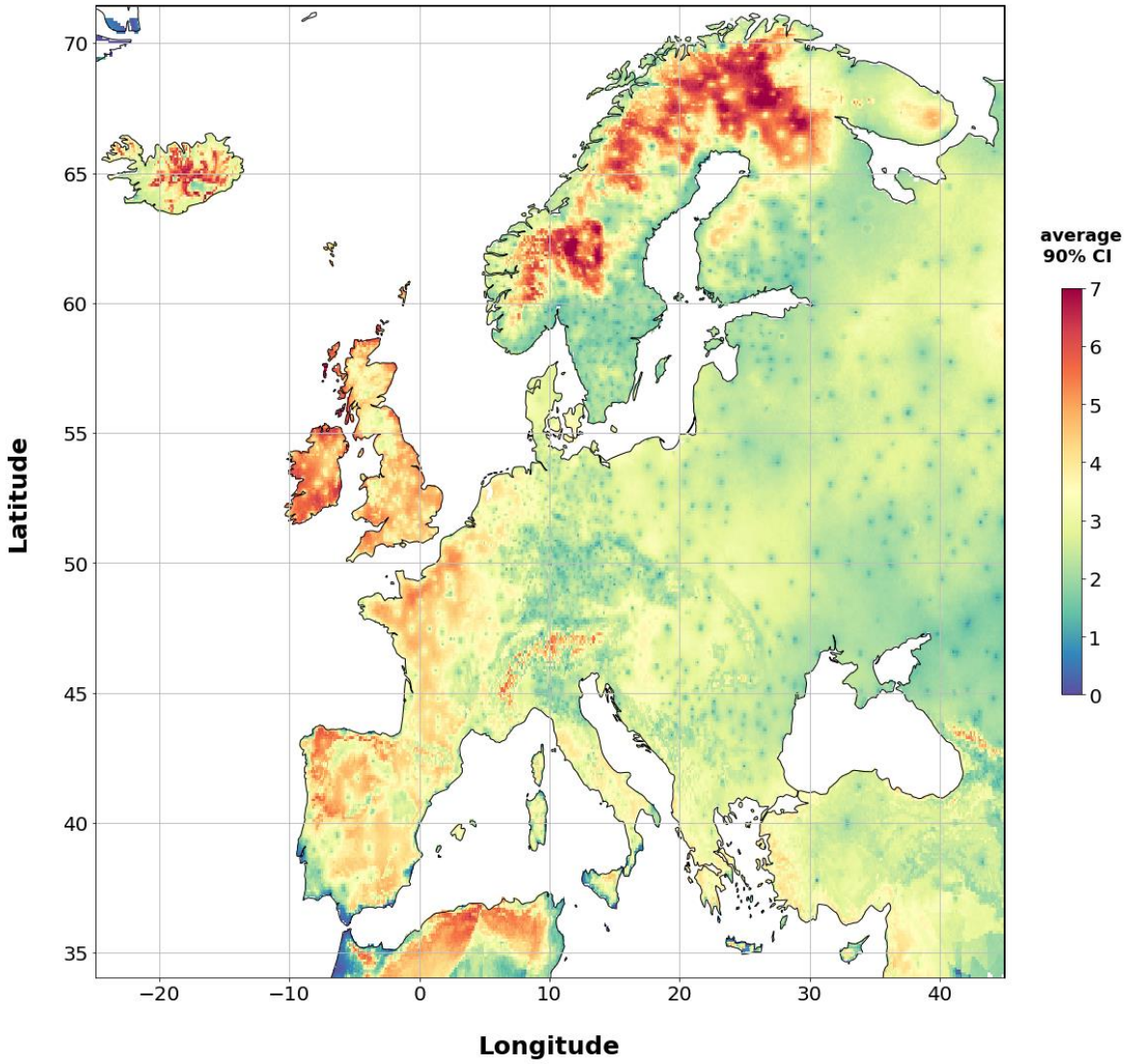
Appendix C. The linear regression DI slope values for 1950 until 2019. The red color indicates positive slopes, which means an increase in DI values, whereas the blue color indicates negative slopes, which means a decrease in DI values. The spatial variability of slope values is relatively high, indicating a strong relationship between location and temporal change of DI. The slope values acquired with linear regression are approximately the same as with the Theil-Sen estimation.

D. Theil-Sen slopes and significance for 1950-1979 and 1980-2019



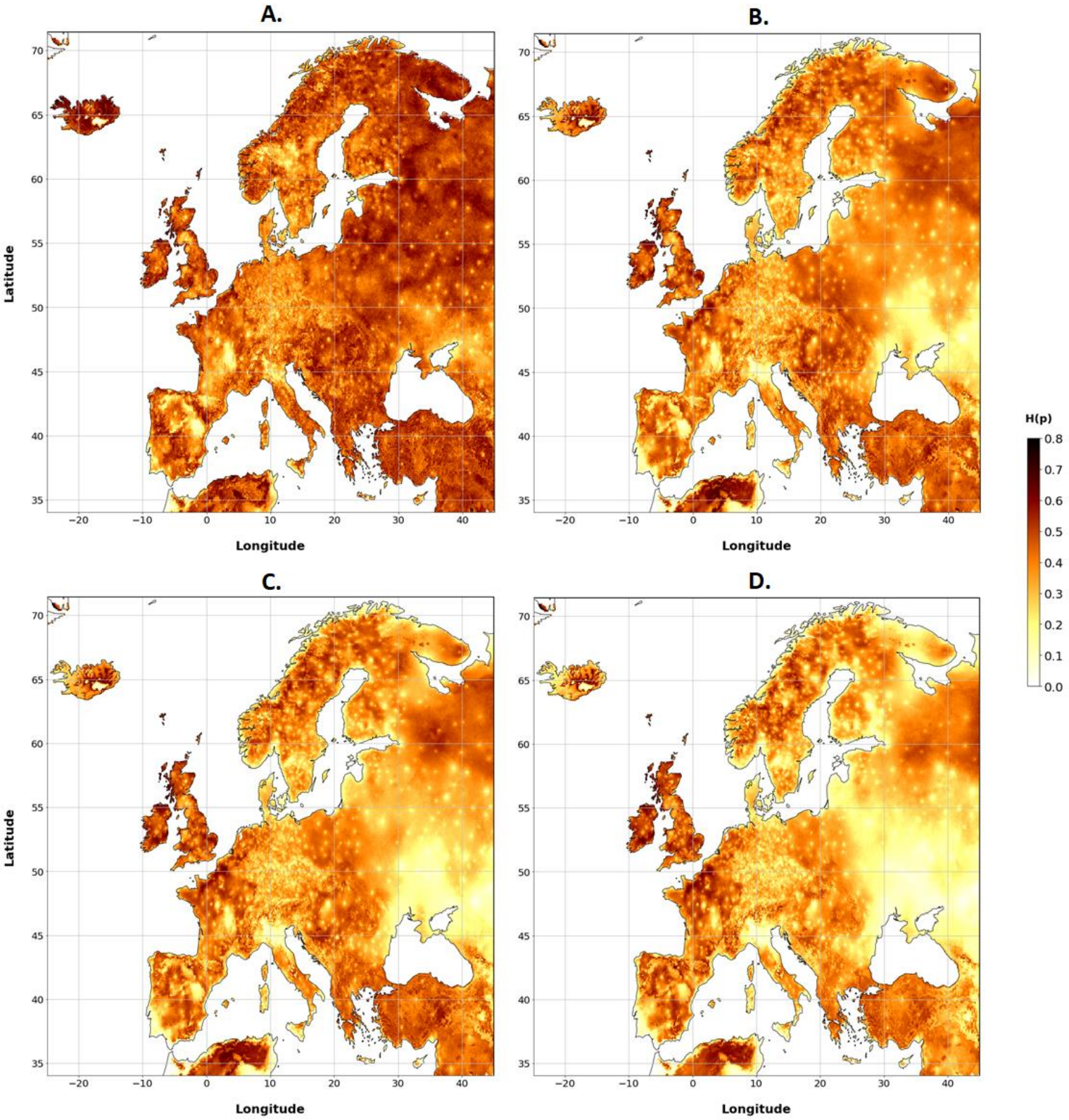
Appendix D. The spatiotemporal patterning of slope and significance values for 1950 until 1979 and from 1980 until 2019. From 1950 until 1980, the slope values are mostly positive, especially in western Europe, whereas the slope values are less prominent and more spatially diverse for 1980 until 2019.

E. Average 90 percent confidence in last freeze date



Appendix F. The average 90 percent uncertainty range for the last freeze date. The last freeze date uncertainty is more influential in determining the false spring uncertainty than the FLD uncertainty, as the last freeze date uncertainty is much higher. Therefore, the patterning of DI uncertainty approximates the patterning that are shown in this figure.

F. Uncertainty of binary false springs



Appendix F. The average uncertainty of (A) 0-, (B) 7-, (C) 10-, and (D) 14-day lag binary false springs calculated with the Shannon's binary entropy function. The uncertainty of binary false springs in general decrease with increasing lag times for areas where the probability of early false springs is already low, for instance in the East European Plains and most coastal areas. Contrarily, the uncertainty of binary false springs generally increases with increasing lag times for areas where the probability of early false springs is high, such as in France and the Spanish Meseta.

G. Python/Dask implementation

A digital version of the model is also available on the following link:
https://github.com/gurensch/Dask_SIx

The GitHub page contains a documentation on how to set-up the model.