Biogeosciences Discuss., 12, C8730–C8744, 2015
www.biogeosciences-discuss.net/12/C8730/2015/

Biogeosciences Discussions

Interactive comment on "Modelling anomalies in the spring and autumn land surface phenology of the European forest" *by* V. F. Rodriguez-Galiano et al.

V. F. Rodriguez-Galiano et al.

vrgaliano@gmail.com

Received and published: 29 December 2015

Anonymous Referee #2 Received and published: 3 November 2015

General comments 1. This ms offers an account of using random forests to characterize land surface phenology (LSP) anomalies over "the European Forest". Unfortunately, the authors do not go further to define the spatial domain of their analysis and there are no maps either in the main article nor in the supplement. This lack of spatial information about the domain is disconcerting, but not as disturbing as the complete lack of discussion or visual display of spatial residuals for a method that the authors repeatedly tell us is able to cope with spatial nonstationarity. When describing the modeling

phenomena in space and time, it is incumbent upon the authors to show, not just tell, the readers about model performance. Aspatial, atemporal goodness of fit metrics, like the pseudo coefficient of determination tells us about just one aspect of the model – its overall ability to explain the appearances. But how well it explains in a particular location or specific period will depend on the selected random forest models. That we should expect the model fits equally well everywhere would be to overlook the purported strength of the modeling approach – its flexibility. That the authors need to map out the model results and analyze the spatial and temporal patterns of residuals; otherwise the reader is left to conclude that the model approach is not as good as the hype.

Reply: A figure showing the spatial domain of the European forest has been included in the manuscript. Additionally, the spatio-temporal distribution of the LSP relative timing values considered in this study is now given in Figures S1 and S2 in the supporting information. The spatio-temporal distribution of the residuals (relative error) computed from the independent test used for Figure 5, is now given as Figures S3 to S6 in the supporting information. However, we would like to clarify that when we described Random Forest in the abstract as a "spatially and temporally non-stationary machine learning approach", our intention was to point out the ability of RF to fit many different regression models within the same overall model. Random Forest, as an ensemble of Regression Trees, is able to learn different temporal and spatial patterns at each node, integrating them into a unique overall model. This property of RF makes it potentially useful for studies with a high spatial and temporal variation in the relationships between the target variable and the predictors. We would like to reiterate that the objective of the paper is not to forecast LSP, but rather to fit a predictive model between climate variables and temporal variation in LSP. RMSE values are given for the RF and Linear Regression models shown in Figure 5, and for those RF models shown in the tables of the Supporting information.

2. The authors have chosen to focus on modeling LSP anomalies rather than the

LSPs directly. OK, that is fine to explore potential drivers of LSP variation and change, but the authors have chosen to produce the anomalies in a non-standard way that shrouds the underlying linear decomposition idea of anomalies. The authors state the follow- ing on 11839 l25f: “The value of the targeted year was excluded in the computation to enhance the interannual variation.” While the authors cite Saleska et al. 2007 as precedent, that doesn’t mean such exclusion was a good idea. First, the long-term (here 11 years) average is now contingent depending on year, so there are 11 long-term averages, each based on 10 years. Furthermore, if the targeted year was excluded in the calculation of the average, then it should have also been excluded in the calculation of the standard deviation, so now there are also several standard deviations. Whether 10 or 11, there are too few years to get a stable estimation of the standard deviation and thus the significance of the various anomalies are called into question. Finally, it is indeed difficult to reconstitute the observational field from the anomaly field due to the exclusion of the targeted years. A simple temporal median and its deviation would be a better approach to maintain decomposability and transparency.

Reply: We are aware that climate scientists generally choose the base as a moving 30 year period and study the changes with respect to that. From a remote sensing perspective, the only dataset that cover a temporal period of thirty years is the AVHRR GIMMS NDVI time series (July 1981 to December 2011). However, its spatial resolution can be too coarse (8 km), making it impossible to distinguish between different land covers (i.e. forest). This factor is of paramount importance in identifying the main drivers of changes in forest LSP, as changes between crops and other land covers such as grasslands must be filtered out. This relates not only to the fact that different land covers can be controlled by different drivers, but also to the fact that croplands are usually irrigated in Europe, so the impact of natural drivers such as temperature or precipitation could be confounded at this spatial resolution. This is why we have decided to use a dataset at 1 km spatial resolution, even though the duration of the time-series is limited to 10 years. We agree with the reviewer that a period of 10 years is short for the rigorous detection of anomalies. However, this is expressly not

the purpose of this manuscript. Anomaly computation is used in this manuscript as a proxy to measure temporal variation in LSP within a decade. To make this emphasis absolutely clear, we have changed the title of the manuscript to “Modelling temporal variation in the spring and autumn land surface phenology of the European forest” and updated all the references to it throughout the manuscript. Additionally, the paragraph explaining the approach followed for computing temporal variation has been rewritten as follows: “LSP parameter estimates (i.e., timing of LSP) during the study period were standardised such as to focus on the relative timing of LSP, rather than absolute timing of LSP. This had the advantage that spatial variation between pixels in the (absolute) timing of LSP was normalised out, leaving only information on whether the LSP parameters were relatively early or late in a given year. The standardised normal deviate (z-score) values for a given pixel can be defined as the differences from the multi-year mean, normalized by the standard deviation across years. Here, the value of the targeted year was excluded in the computation of the multi-year mean to enhance the inter-annual variation (Saleska et al., 2007). To retain and convey a clear sense of the present meaning, we refer to this standardised variable throughout as “LSP relative timing”. The spatio-temporal distribution of spring and autumn LSP relative timing values is shown in Figures S1 and S2 of the supporting information, respectively.”

3. Perhaps more important to the intended message of this ms is that modeling LSP anomalies is an activity distinct from modeling LSPs. What is the object of the modeling? The former explains the appearances to gain insight in what kinds of things drive deviations from a simple temporal average. The latter aims both to explain the past temporal patterns and to provide a means for forecasting LSP. This former object is what the authors try to do by sifting through variable importance. But note, the authors do not explore in this ms the spatially explicit pattern of variables. Instead, all that is provided are global evaluations of the relative importance of each variable to explaining the overall variation of the dataset. Most of the literature on LSP seeks to do the latter and thus it is misleading or at least inaccurate to cite papers doing the latter as inadequately doing the former. Different objectives requires different methods.

Reply: The objective of this manuscript is defined at the end of the introduction as follows: “The aim of this study is, therefore, to provide an explanation of the observed temporal variation in LSP of the entire European forest during the last decade, identifying the main weather drivers for spring and autumn at the continental scale. Our research offers new insights into the study of LSP by modelling the climate-driven past temporal variation in phenology, rather than trends, and using innovative multivariate non-linear machine learning techniques to evaluate multiple weather predictors at biological scales, and non-weather predictors such as the legacy effect of the date of spring onset in leaf senescence. Climate predictors used range from 30 days average values of temperature variables (max, min and avg) such as precipitation, short wave radiation and day length; trimestral cumulated values such as growing degree days or chilling requirements, among others; to the date of specific events such as the first freeze or the last freeze. Moreover, we considered flexible biological time scales in the analysis between weather and phenological events rather than calendar months.” We want to be absolutely clear that our objective is therefore the first option, as we are trying to explain temporal variation in phenology and forecasting is not the objective. We believe that the objective of the paper was clearly stated in the first version of the manuscript. However, it is true that perhaps a couple of sentences in the introduction might be confusing and might lead to misunderstandings or decontextualize some of the references given. In particular the following sentences: “This last group of studies focuses on changes in phenology produced by average changes in weather (mainly trends in warming). However, changes in LSP arising as a consequence of the inter-annual variability in weather are relatively unstudied, with model-based studies of this phenomenon being scarce (van Vliet, 2010).” By average changes we meant trends and anomalies was a synonym of temporal variation in LSP. In order to make it clearer this paragraph has been rewritten as follows: “Although different approaches have been devised for the study of vegetation phenology (Rafferty et al., 2013), the characterisation and modelling of vegetation phenology at global or regional scales has been undertaken mainly through the use of long-term time-series of satellite-sensor vege-

tation indices (termed land surface phenology, LSP, to reflect that satellite-observed phenology includes all land covers). Most studies of LSP analyse trends in phenological events across years (Delbart et al., 2008; Jeganathan et al., 2014; Jeong et al., 2011; Karlsen et al., 2007; Myneni et al., 1997), but more recent studies present process-based models to uncover cause-effect relationships between long-term trends in phenology and its key driving variables (Ivits et al., 012; Maignan et al., 2008a; Maignan et al., 2008b; Stöckli et al., 2011; Stöckli et al., 2008; Yu et al., 2015; Zhou et al., 2001). This last group of studies focuses on trends in phenology produced by trends in weather (mainly warming). However, temporal variation in LSP arising as a consequence of the inter-annual variability in weather are less studied (Cook et al., 2005; De Beurs and Henebry, 2008; Menzel et al., 2005; Post and Stenseth, 1999; Zhang et al., 2004), with model-based studies of this phenomenon being scarce (van Vliet, 2010).”

4. While the authors’ enthusiasm for RF is understandable, their line of argument frequently devolves into boosterism rather than a serious comparison of linear methods to modeling LSP or LSP anomalies. As an apparent afterthought, in the Discussion section starting on p 11847 l 12f, there is a mention that the authors also did multivariate linear regression on the same variables and achieved coefficients of determination of 0.36 and 0.26 for spring and fall anomalies across the spatial domain. Unfortunately the details of what they did are too few and there are no direct comparisons spatially or temporally so it is hard to know just how poorly a “typical” modeling approach worked. The comparison of the predicted versus observed phenology anomalies for RF and the ill-described linear regression models (MLR) relies on highly leveraged coefficients of determination rather than RMSEs and biases.

Reply: In order to provide a clear comparison with linear methods, a performance matrix for Random Forest and multivariate linear regressions (MLR) has been added in the current version of the manuscript: âĂć RMSE values are included for the independent test shown in Figure 5 (Figure 4 of the previous submission) for both RF and MLR. âĂć The spatio-temporal distribution of the residuals (RF and MLR) computed from the

independent test used for Figure 5, is now given in Figures S3 to S6 in the supporting information. âĂć Details on the MLR models are given in Table S4.

5. Certainly there is not sufficient evidence presented in the ms to conclude as stated in the end of the abstract: “This research, thus, shows clearly the inadequacy of the hitherto applied linear regression approaches for modeling LSP and paves the way for a new set of scientific investigation based on machine learning methods.” Hardly! If the authors want to achieve that end, then they need to construct a very different kind of inter-comparison study.

Reply: We agree with the referee that we might have fallen into an excess of enthusiasm by writing this. We have therefore rewritten the sentence as follows: “This research, thus, shows an alternative to the hitherto applied linear regression approaches for modelling LSP and paves the way for further scientific investigation based on machine learning methods.”

Specific comments P 11835 l 13 what was the precision of the spring index method used in the Schwartz et al. 2006 paper? What was the precision of the extended Spring Indices published in papers earlier this year by Schwartz and his colleagues? The SI and SI-x use daily meteorological data as input.

Reply: We are aware about the precision of the spring index, and of the paper of Yu et al. published a few months ago in Int. J. Biomeorol. Actually, both references are cited in our manuscript on numerous occasions. However, we do not understand why we should refer to this paper in our abstract as P 11835 l 13 refers to the abstract . Yu et al. (2015) showed that a multivariate linear model using GDD and photoperiod could explain 87.9 to 93.4% of the variation in spring canopy development and 75.8-89.1% of the variation in autumn senescence. These are very good results, of course, and we recognise the quality of spring indices. However, the performance of those of Yu et al. and our models cannot be compared in absolute terms, for several reasons: 1. Yu et al. performed their analyses using ground phenological observations of 106 individual

trees in a temperate deciduous woodlot at Milwaukee. We are modelling LSP relative timing values at the Pan-European continent scale. It is not only that the scale of the study is different; it is also that LSP is affected by numerous uncertainties that make its modelling extremely challenging (see Rodriguez-Galiano et al 2015 in GRL). 2. Yu et al. used air temperature recorded every 10 min by 4 temperature sensors connected to data loggers and located in an area of 4.5 ha. We are using grids of interpolated ground stations at a continental scale. These dataset are affected by numerous uncertainties as well. For instance, the resolution of the grids is equal to 0.25 degrees, the spatial distribution of meteorological stations is variable and not all the pixel′s values might be representative of the actual weather for all the forest patches. 3. Yu et al. were modelling trends in phenology, whereas we are modelling temporal variation. However, we agree with the referee that studies on changes in phenology are scarce, so it might be necessary to compare with studies focused on trends.

p 11835 l 12 are these relative errors referring to the day of the year?

Reply: No, the relative errors referred to a percentage of the observed LSP relative timing values (z-score value).

P 11836 l 4 no, not to reflect all land covers, which are a human concept, but rather that remote sensing of phenology relies fundamentally on reflected shortwave radiation upwelling from mixtures of things at the surface. Moreover there are no well-defined phenophases for land covers as there are for specific species, and we have well defined phenophases for on a fraction of the dominant or economically important plant species.

P 11836 l 12 “relatively unstudied” is hardly accurate. There is a robust line of papers looking at climate mode influence on LSPs. Some older examples include: Tucker et al. 2001 IJBM, Buermann et al. 2013 JGR, Gong & Shi 2003 IJRS, Potter et al. 2003 GCB, Zhang et al. 2004 GCB, Cook et al. 2005 GCB, de Beurs & Henebry 2008 JClim. And the field phenology community has long known this as well, e.g., Post & Stenseth 1999 Ecology, Menzel et al. 2005 GCB.

Reply: We agree with the reviewer that this is an oversight. We have now replaced "relatively unstudied" by less studied and included some of the cites proposed by the reviewers.

P 11837 l 15f testing for a legacy effect in the LSPs is an interesting idea, but scaling issues would suggest that any signal would be swamped.

Reply: Yes, we agree at the scale of study it would be difficult to identify any legacy effect. However, we wanted to see if the legacy effect can be characterised at a landscape scale.

P 11838 l 17 by focusing on "climate-driven anomalies in phenology" changes in LSPs that arise from disturbances such as fire or flooding and changes in land cover or land use or harvesting in agroforestry plantations will all be overlooked and simple contribute to the noise in the system.

Reply: This study is at the continental scale, and we agree that some variation in phenology are not climate-driven, but related to other natural and anthropogenic factors. This is one of the reasons that we decided to apply the Random Forest. RF is able to handle noisy datasets because it is an ensemble of trees and every tree is trained from different subsets of data (this is intended to increase diversity of patterns and to reduce the effect of noise). On the other hand, only LSP estimates with complete temporal coverage were included and when the LSP relative timing values were resampled from 1 km to 25 km (approximately), the median of at least 50 estimates was computed. This should exclude most of the disturbances (or 'noise') which mostly occur at a smaller scale. Please, see the following paragraph in the methods section: "To match the spatial resolution of the ECA&D dataset, the LSP relative timing values for each year were resampled to a spatial resolution of $0.25^\circ \times 0.25^\circ$ by calculating the median of all the LSP relative timing values within this area after excluding the areas with fewer than 50 LSP estimates and the non-forest pixels according to the Globcover2005 and Globcover2009 land cover maps (http://due.esrin.esa.int/globcover/). Only LSP

estimates with complete temporal coverage (2003–2011) were included in the analysis to reduce the likelihood of natural and human disturbances (Potter et al., 2003). Globcover was selected for its greater consistency with the MERIS MTCI time-series and its high geolocational accuracy (<150 m) (Bicheron et al., 2011)."

P 11838 l 21 "monthly average values"!? Below it is specified that daily data were used. Which is it?

Reply: Yes, daily data were used in our analysis. However, the average of the 30 days prior to the phenological events is one of the functions that was implemented to compute the weather predictors. "Monthly average values" has been replaced by "30 days average values" to avoid any misunderstanding.

P 11838 l 25f "rather than fixed calendar dates" this is a strawman argument. Who uses "fixed calendar dates"?

Reply: "Fixed calendar dates" has been replaced by "calendar months".

P 11839 what are the units of measurement in the DAL and SIS?

Reply: The units for both of them are w m-2. This has been included in the manuscript.
P 11839 l 20-21 "using the methodology described in Dash et al. (2010)." Succinctly describe the method.

Reply: We have included a brief description of the method: "The time-series of MERIS MTCI data was used to estimate both the onset of greenness (OG) and end of senescence (EOS) from 2003 to 2011. Data for every estimation year considered 1.5 years of data (from October in the previous year to July in the next year) because the annual pattern of vegetation growth in some parts of Europe spans across calendar years and, hence, insufficient information about LSP is captured using a single year of data. The yearly values of OG and EOS were estimated for each image pixel of the study area using the methodology described in Dash et al. (2010). This methodology consists of two major procedures: data smoothing and LSP estimation (Figure 2a). Smoothed

MTCI time-series data were obtained using a discrete Fourier transform because of its advantage of requiring fewer user-defined parameters compared to other methods (Atkinson et al., 2012). The peak in the annual profile was defined as a point on the phenological curve where the first derivative changes sign from positive to negative. Next, the derived data were searched backward and forward departing from the maximum annual peak to estimate the OG and EOS, respectively. OG was defined as a valley at the beginning of the growing season point (a change in derivative value from positive to negative) and EOS was defined as a valley point occurring at the decaying end of a phenology cycle (a change in derivative value from negative to positive). These satellite-derived LSP estimates were compared to ground observations of the thousands of deciduous tree phenology records of the Pan European Phenology network (PEP725) (Rodriguez-Galiano et al., 2015a). This comparison resulted in a large spatio-temporal correlation of the phenology estimates with the spring phenophase (OG vs leaf undolding; pseudo-R2=0.70) and autumn phenophase (EOS vs autumnal colouring; pseudo-R2=0.71)."

P 11839 l 25 "long-term mean" is a stretch for 11 years of data. How about using "multi-year mean", if the suggested approach of the median is not adopted.

Reply: The term "long-term" has been replaced by "multi-year".

P 11840 l 2 resampling the 1km and 0.05 degree data to 0.25 degrees will greatly degrade the LSP signal. Perhaps a map of the residuals from the median will help to understand the spatial performance of the RFs.

Reply: The residuals are given now in the Supporting information.

P 11840 l 4 change "with less than 50" to "with fewer than 50" since the quantity is countable.

Reply: This has been corrected.

P 11840 eschew "Julian date" and use "day of year (DOY)" instead.

Reply: This has been implemented . P 11840 l 17f Clarify: "Relative differences of the climatologies. . .." What climatologies? P 11840 l 21f put this information with units into a table

Reply: A table showing all the predictors used has been included in this version of the manuscript

-See Table 1 in the manuscript (attached compressed file)

P 11841 l 13 "and may vary spatially" but not temporally? P 11842 15 decode "oob"

Reply: This has been corrected, "spatially and temporally". "oob" has been expanded.

P 11842 l 17 "from 1 to 9" but figures 2 & 3 show 12 and 14 predictors

Reply: Yes, this is correct. It is the number of random predictors that are considered at the splitting of each node. It must not be the maximum number of variables. Usually, the best performance is achieved at around 1/3 of the total number of predictors. Considering all the predictors dramatically increases the computation time without any increase in the model′s accuracy.

P 11842 l 25-26 "rather than human-imposed temporal scales" but assigning 30 or 90 day windows is a "human-imposed" scale, it is not? , P 11844 l 3 "absolute fixed dates of the calendar months" there is that strawman argument again!

Reply: Yes the number of days is imposed, but not the beginning of the period. "Absolute fixed" dates has been replaced by "calendar months"

P 11845 l 16 "climatic predictors" the predictors used relate to weather, not climate.

Reply: This has been replaced throughout the manuscript.

P 11845 l 25 "improving the temporal matching between LSP anomalies and the preceding climatic anomalies" the ms does not show this.

Reply: This is the updated statement written in the current version of the manuscript:

"Our study advanced the modelling of vegetation phenology by improving the temporal matching between LSP temporal variation and the preceding weather conditions by analysing daily data at biological scales."

P 11846 l 3 "a consistent divergent effect was observed between spring and autumn" what is this? Not clear. The order of the sentences has been altered to improve interpretability:

Reply: "Regarding, the length of the temporal windows for weather function computation, Menzel et al. (2006) showed that most phenological phases of plant species in Europe correlate significantly with mean temperatures of the month of onset and the two preceding months. However, in our study, when end of senescence was considered, a consistent divergent effect was observed between spring and autumn. Autumn phenophases might be driven by longer-term changes in weather, while for spring the average conditions of the 30 days previous to the date of onset play a more important role (Table S1, S2 and S3 in supplementary information). From a computational point of view, considering larger temporal windows for calculating averages would induce a smoothing effect, degrading the information in the predictors, whereas cumulative functions such as GDD or chilling requirements would not be affected by this effect. However, we observed a divergent response between spring and autumn and consistent throughout the models of each phenophase suggests that a biological explanation for this phenomenon might be plausible."

P 11846 ll10-12 that spring is in the mid to high latitudes is driven by the timing of the thaw is not new information, nor is the observation that the arrival of autumn is a complicated, contingent process depending on the growing season's trajectory coming into the fall season. We agree that the major drivers are known or mentioned earlier, but we demonstrated application of an non-linear modelling approach at a continental scale with a better ability to predict changes in phenology compared to standard linear models. We also demonstrated the combination of drivers in initiative key phenological events rather than single driver. P 11846 l 29 "our results support this hypothesis" what

hypothesis exactly? Restate it here succinctly.

Reply: This has been rewritten as follows: "Our results reveal that multiple environmental drivers are required to initiate phenological events of Europe and also showed that the role of GDD alone in driving spring phenology might be overestimated due to an over-reliance on linear models."

Figure 1 has 3 typographical errors in the spelling of anomalies and explanative should be explanatory. Where is the legend to decode the significance of the shapes?

Reply: This figure has been modified according to both reviewer′s comments. See Figure 2 in the manuscript (attached compressed folder).

Figure 2 what is the scale of "relative error"? should it read 0-200%?

Reply: Yes, it is 200%. It is kept to this magnitude because we wanted to show the performance of all the possible models regarding the composition of predictors. It should be noted that these values are only reached by the models of a very poor performance. There are also some results from the models with a good performance (just a few) that can reach high values of relative errors. These estimates might be associated to outliers. It should be noted that this test includes thousands of samples. However, we preferred not to filter the data (training and testing) to avoid introducing a bias into the analysis.

Figure 3 has several shortcomings. First is the poor contrast in pseudo-r2. Move these important numbers to the top of the bars. (Note that r2 is called the coefficient of determination, not the "square correlation coefficient".) Second, it is not clear why there are 12 and 14 climatic drivers shown when the pseudo-r2 is stable after 8 or fewer. How to determine when the model is sufficiently rich?

Reply: The figure has been modified according to the referee′s comments. All the possible models regarding the number of weather predictors are shown. We believe that it is interesting to show all the models because in this way it is possible to see

the differences between using just one predictor (i.e. temperature) or using multiple predictors. The model was selected based on the coefficient of determination and the relative errors. As can be seen, this is only important for the accuracy of the estimates, but it does not affect the ranking/importance of the weather drivers.

Figure 4 is a classic case of a highly leveraged regressions. Not one of these patterns show good performance. What are the biases and RMSEs of these models? Relative errors are shown in the previous figure. RMSE values are given now in the figure caption and the temporal and spatial distribution of the residuals is shown in the supporting information.

Reply: We consider that the performance of the RF model was satisfactory (upper panel), unlike linear regression models. The explained variances for spring and autumn RF models are 90% and 0.43 (spring Random Forest model) and 68% and 0.92 (autumn Random Forest model), respectively. However, we would like to highlight the complexity of the problem that we are trying to model (see previous comments). Cook et al. (2005 in GCB) in a similar study explained 63% of the variance in onset for mixed and boreal forest in Europe using GDD only. RMSE values obtained from Cook et al. are not reported.

Please also note the supplement to this comment:
http://www.biogeosciences-discuss.net/12/C8730/2015/bgd-12-C8730-2015-supplement.zip

---

Interactive comment on Biogeosciences Discuss., 12, 11833, 2015.