Ecosystem Insights through Extreme Values: A Fresh Look at Meteorological Drivers

Christian Reimers, Claire Robin and Alexander Winkler
Department of Biogeochemical Integration
Max Planck Institute for Biogeochemistry
Jena, Germany
creimers@bgc-jena.mpg.de

Abstract

Understanding the influence of meteorological drivers on the ecosystem is a central problem in Earth's system science. Deriving these influences directly from observations is crucial. However, natural systems often have complex interactions, where multiple drivers influence each other and the target variable simultaneously. These interactions complicate our understanding of individual variable effects. For instance, the relative importance of soil moisture and temperature on net biome production remains unclear, with in-situ measurements and earth's system models yielding contradicting results. In this work, we propose a novel approach: training on extreme values only. By focusing only on these values, we can better approximate the influences of meteorological drivers. We demonstrate the potential of this approach through a simple example, validating it analytically and empirically.

1 Introduction

A major source of uncertainty in climate projections is the reaction of vegetation to increased atmospheric CO_2 concentrations and changes in climate [2]. A pathway to reduce this uncertainty is to understand the relative importance of drivers of ecosystem productivity, the amount of CO_2 the ecosystem will take up. Two of the most important drivers of ecosystem productivity are water availability and temperature. However, determining the relative importance of these drivers is challenging. Multiple papers have come to different conclusions concerning the relative importance of temperature and soil moisture on ecosystem productivity. For example, [3] train a decision tree-based model on in-situ eddy covariance measurements. They find that while soil moisture anomalies are most important to explain the variation in net biome productivity (NBP) anomalies locally, these effects cancel out globally such that the global NBP anomalies are mostly driven by the global temperature anomalies. In contrast, [6] run multiple large Earth system models and find that suppressing the anomalies in soil moisture in these models also suppresses more than eighty percent of the global NBP anomalies, indicating that, in these models, the effects of soil moisture anomalies do not cancel out. One possible explanation for this difference is the interaction between these drivers, as discussed in [1]. The authors evaluate the interaction between temperature and soil moisture through evaporative cooling. They show that on a global scale, suppressing soil moisture anomalies also leads to a suppression of temperature anomalies, indicating that there is a strong indirect effect in addition to the direct effect. They conclude that the difference observed by [3] and [6] are due to attributing this indirect effect to either temperature or soil moisture. However, the work of [1] cannot fully close the gap between the works of [3] and [6]. For example, when attributing the indirect effect to soil moisture, also the local importance of temperature vanishes and only soil moisture is important, which is not consistent with the results from in-situ measurements. In this work, we present an avenue to better learn the direct effect of drivers, training only on extreme values. We motivate this idea from a causal modeling perspective and take a simplified model of

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

atmosphere-biosphere coupling in which we can prove analytically that this approach approximates the true impact of different drivers while training on average data does not. Additionally, we simulate data following this model and show that the better approximation outweighs the worse signal-to-noise ratio that arises from only training on a fraction of the data.

2 A Motivation from a Causality Persepective

In Earth's system science, causal graphs are a useful tool for understanding complex relationships between variables [5]. A causal graph is a graph in which variables are represented as nodes, and directed edges between nodes indicate the direction of causal influence. In other words, an edge $A \to B$ indicates that intervening on A has a "direct" effect on B, meaning that if we change A while keeping all variables except A and B constant, B will also change. In these graphs, many calculations are infeasible. To be able to make inference, we have to decide which connections are significant and which connections can be neglected. The graph might, however, remain complex after removing insignificant links and calculations might yield results that are strongly influenced by noise. The threshold of what is negligible depends on the strength of an effect in addition to the level of noise. During extreme events, we normally find that the influence of one (or a group of) variables is very strong, which raises the threshold as the relative influence of other processes shrinks. This leads to an even sparser graph, where some of the processes might be easier to extract.

3 A Simple Example

We show the advantage of training on extreme values in a simple example: In this example, the relationship between net biome production (NBP), temperature (T), and soil moisture (SM) is

$$NBP = \alpha T_r + \beta SM_r + \gamma + \varepsilon. \tag{1}$$

Here, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is a noise term representing the variation that is not determined by temperature or soil moisture. Further, we assume that we are not measuring the temperature at the leaf but in the air some distance away from the canopy, and the soil moisture is not measured at the root of the plant but some distance away from it. Therefore, the temperature and soil moisture interact through evaporative cooling before they influence NBP. To this end, we assume that the two values (T_r, SM_r) that directly influence NBP can be calculated from the measured values T_{in} and SM_{in}. The amount of evaporative cooling depends on the product ($T_{in}SM_{in}$) and is limited by a maximum amount that can cool/evaporate. The intuition is that only a combination of high temperature and high soil moisture can lead to strong evapotranspiration and, that the maximum effect of evapotranspiration becomes limited by the conductance of the surface. Therefore, we model T_r and SM_r as

$$T_r = T_{in} - \frac{C_T}{1 + e^{-C_1(SM_{in}T_{in} - C_0)}} \quad \text{and} \quad SM_r = SM_{in} - \frac{C_{SM}}{1 + e^{-C_1(SM_{in}T_{in} - C_0)}}. \tag{2}$$

Depending on whether the values are extreme (SM_{in} and T_{in} large), we can make different simplifying approximations: For non-extreme values, we can use the fact that $SM_{in}T_{in}$ is close to the critical constant C_0 and approximate the exponential interaction with a first-order Taylor approximation. For extreme values, we use the fact that $SM_{in}T_{in}$ is much larger than the critical constant C_0 and hence, the exponential interaction is at its maximum. Further, we assume that the interaction is still small compared to the variations in the individual variables and, therefore, can be approximated by two univariate regressions. Using these approximations, we arrive at the following corollary.

Corollary 1. *If we can make the above assumptions, a linear regression on T_{in} and SM_{in} does not return the correct values α and β but*

$$\alpha^* = \frac{\alpha}{1 - C_T SM_{in} C_1} \quad \textit{and} \quad \beta^* = \frac{\beta}{1 - C_{SM} T_{in} C_1}. \tag{3}$$

However, if we train only on extreme values, we find $\alpha^* = \alpha$ and $\beta^* = \beta$.

The calculations can be found in Appendix A. This result shows that when using linear regression to estimate the importance of temperature and soil moisture on NBP, training on extreme values provides more accurate estimates, whereas training on non-extreme values can lead to biased estimates of the individual importances due to the interaction between the variables.

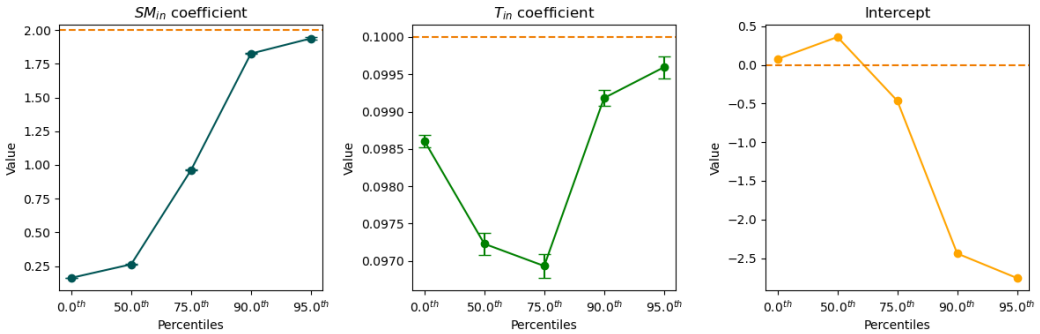

Figure 1: Plots of the estimated importance of the drivers and the bias when trained on only a percentile of the data. Left: The values of β, the linear coefficient of soil moisture, estimated on only data with values above different percentiles (0, 50, 75, 90, 95) of the data. Middle: The same plot for α, the linear coefficient of temperature. Right: The same plot for the bias of the prediction. The dotted line in each plot represents the correct value (2, 0.1, 0) for the estimated quantity.

4 Emperical Experiments

To demonstrate that our approximations are reasonable, we simulate some data following the setup described in Section 3. Details can be found in Appendix B and results in Figure 1. We find that using only the top 5% of data significantly improves the accuracy of estimating α and β, reducing the error by 96.6% and 85.7%, respectively, compared to using all the data. However, this improvement comes at the cost of increased bias, as the mean difference between predictions and labels increases. Specifically, the error in bias is 96.9% larger when calculated using only the top quantiles compared to using all the data. This trade-off highlights the benefits and limitations of training on extreme values, which can better capture the influence of interacting drivers but may also introduce additional bias.

5 Discussion

In the last three sections, we motivated why training on extreme events should give a better approximation of the importance of different drivers. We further presented some mathematical arguments and empirical simulation results to underpin the effectiveness of this idea. However, this is only a first step, and a lot of future research is needed before we can rely on this method for scientific discovery. First, our simulated example is very idealized. As a next step, we need to test the relation again with a more complex model, like the CLASS model[7]. Further, the system we have simulated has a saturation in the interaction but no saturation in the target variable. While we believe that this is a reasonable model for the interaction of temperature, atmospheric vapour pressure, and soil water content, it is unclear whether this can also be applied to other situations in Earth's system science. Additionally, for this study, we only use a linear model. While it is common to linearize effects in order to calculate the importance of a driver, it would be good to understand whether the effect also holds if we use a neural network instead. Finally, it would be a good idea to test the difference between calculating the importance of different drivers on an observational dataset, for example, the FLUXNet2015 dataset [4].

6 Conclusions

In this work, we proposed a method to estimate the importance of individual drivers of an ecosystem variable, where drivers interact with each other. We motivated this method using causal inference and demonstrated its effectiveness using in a simple example analytically and empirically. Our results show that training on extreme events improves the accuracy of driver influence estimates but increases the bias in the prediction. This is a first step, and further validation with more complex examples is needed before applying the method. However, it could help resolve the disputed global influence of water and temperature on net biome production.

Acknowledgements

Funding for this study was provided by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187)

References

[1] Vincent Humphrey, Jakob Zscheischler, Philippe Ciais, Lukas Gudmundsson, Stephen Sitch, and Sonia I Seneviratne. Sensitivity of atmospheric co2 growth rate to observed changes in terrestrial water storage. *Nature*, 560(7720):628–631, 2018.

[2] IPCC. Climate change 2023: Synthesis report. contribution of working groups i, ii and iii to the sixth assessment report of the intergovernmental panel on climate change, 2023.

[3] Martin Jung, Markus Reichstein, Christopher R Schwalm, Chris Huntingford, Stephen Sitch, Anders Ahlström, Almut Arneth, Gustau Camps-Valls, Philippe Ciais, Pierre Friedlingstein, et al. Compensatory water effects link yearly global land co2 sink changes to temperature. *Nature*, 541(7638):516–520, 2017.

[4] Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, et al. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):225, 2020.

[5] Judea Pearl. *Causality*. Cambridge university press, 2009.

[6] Sonia I Seneviratne, Micah Wilhelm, Tanja Stanelle, Bart Van Den Hurk, Stefan Hagemann, Alexis Berg, Frederique Cheruy, Matthew E Higgins, Arndt Meier, Victor Brovkin, et al. Impact of soil moisture-climate feedbacks on cmip5 projections: First results from the glace-cmip5 experiment. *Geophysical Research Letters*, 40(19):5212–5217, 2013.

[7] Jordi Vilà-Guerau de Arellano. Atmospheric boundary layer: Integrating air chemistry and land interactions. *Atmospheric Boundary Layer: Integrating Air Chemistry and Land Interactions*, 2015.

A Calculations

In this section, we show the calculation for the values of α^* and β^* presented in Corollary **??**.

For non-extreme values, we can make some simplifications. First, we assume that the normal values are such that the product $SM_{in}T_{in}$ is close to the critical constant C_0. Hence, we can use the first-order Taylor approximation for the exponential:

$$e^x \approx 1 + x \tag{4}$$

which is valid for x close to zero. To use this approximation, we first extend the fraction to

$$T_r = T_{in} - \frac{C_T}{1 + e^{-C_1(SM_{in}T_{in} - C_0)}} = T_{in} - C_T \frac{e^{C_1(SM_{in}T_{in} - C_0)}}{e^{C_1(SM_{in}T_{in} - C_0)} + 1}. \tag{5}$$

Now, we can use the approximation for both parts of the fraction to find

$$T_r \approx T_{in} - C_T \frac{1 + C_1(SM_{in}T_{in} - C_0)}{2 + C_1(SM_{in}T_{in} - C_0)}. \tag{6}$$

By using some simple modifications to the term, we find

$$\begin{aligned} T_r &\approx T_{in} - C_T \left(\frac{2 + C_1(SM_{in}T_{in} - C_0) - 1}{2 + C_1(SM_{in}T_{in} - C_0)} \right) \\ &= T_{in} - C_T \left(1 - \frac{1}{2 + C_1(SM_{in}T_{in} - C_0)} \right) \\ &= T_{in} - 2C_T \left(\frac{1}{2} - \frac{1}{1 + C_1(SM_{in}T_{in} - C_0)/2} \right). \end{aligned} \tag{7}$$

Now, we use the abovementioned approximation in the opposite direction to get

$$T_r \approx T_{in} - 2C_T \left(\frac{1}{2} - \frac{1}{e^{C_1(SM_{in}T_{in} - C_0)/2}} \right) = T_{in} - 2C_T \left(\frac{1}{2} - e^{-C_1(SM_{in}T_{in} - C_0)/2} \right). \tag{8}$$

Finally, we use the approximation again and simplify the relation to

$$\begin{aligned} T_r &\approx T_{in} - 2C_T \left(\frac{1}{2} - \left(1 - \frac{C_1(SM_{in}T_{in} - C_0)}{2} \right) \right) \\ &= T_{in} - 2C_T \left(\frac{1}{2} + \frac{C_1(SM_{in}T_{in} - C_0)}{2} \right) \\ &= (1 - C_T SM_{in} C_1) T_{in} - (1 - C_0 C_1) C_T. \end{aligned} \tag{9}$$

Equivalently, we find

$$SM_r \approx (1 - C_{SM} T_{in} C_1) SM_{in} - (1 - C_0 C_1) C_{SM}. \tag{10}$$

While the factors obviously depend on the values of T_{in} and SM_{in}, we can approximate them with constant factors leading to

$$T_{in} = \delta_T T_r + \eta_T \tag{11}$$

and

$$SM_{in} = \delta_{SM} SM_r + \eta_{SM}. \tag{12}$$

If we want to calculate α^* and from the measurements, we assume that the correlation through the interaction is that much smaller than the variation within the variable such that we can use a univariate regression instead of a multivariate regression, we find

$$\begin{aligned} \alpha^* &= \frac{\langle NBP - \mathbb{E}(NBP), T_{in} - \mathbb{E}(T_{in}) \rangle}{\langle T_{in}, T_{in} \rangle} \\ &= \frac{\langle \alpha(T_r - \mathbb{E}(T_r)) + \beta(SM_r - \mathbb{E}(SM_r)), \delta_T(T_r - \mathbb{E}(T_r)) \rangle}{\langle \delta_T(T_r - \mathbb{E}(T_r)), \delta_T(T_r - \mathbb{E}(T_r)) \rangle} \\ &= \frac{\alpha \delta_T \langle T_r - \mathbb{E}(T_r), T_r - \mathbb{E}(T_r) \rangle}{\delta_T^2 \langle T_r - \mathbb{E}(T_r), T_r - \mathbb{E}(T_r) \rangle} = \frac{\alpha}{\delta_T} \end{aligned} \tag{13}$$

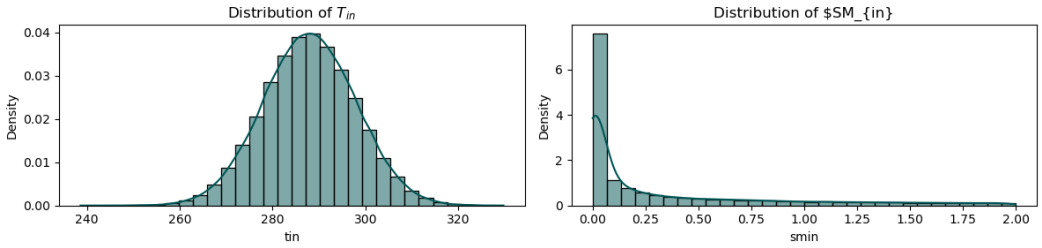

Figure 2: The distributions of simulated temperature and soil moisture. As a first step towards realism, we simulated temperature as a normal distribution centered at 288 K with a standard deviation of 10 K and soil moisture as a very skewed distribution which takes only positive values and is small most of the time.

and analogously

$$\beta^* = \frac{\beta}{\delta_{SM}}. \tag{14}$$

Here $\langle A, B\rangle$ denotes the covariance between A and B.

This means that the α^* and β^* values estimated on this data do not agree with the values from the data creation process.

For high values, meaning $T_{in}SM_{in} >> C_0$ we can use the approximation

$$e^{-(T_{in}SM_{in}-C_0)} \approx 0. \tag{15}$$

This leads to

$$T_r = T_{in} - \frac{C_t}{1+e^{-(T_{in}SM_{in}-C_0)}} \approx T_{in} - C_t. \tag{16}$$

and analogously

$$SM_r = SM_{in} - C_{SM}. \tag{17}$$

Now, calculating the linear regression between the observations of T_{in}, SM_{in} and NBP where we again use the assumption that the interaction is so much weaker than the variance in the variables that we can use a single variate regression instead of a multivariate regression, we find

$$\begin{aligned}
\alpha^\dagger &= \frac{\langle NBP - \mathbb{E}(NBP), T_{in} - \mathbb{E}(T_{in})\rangle}{\langle T_{in} - \mathbb{E}(T_{in}), T_{in} - \mathbb{E}(T_{in})\rangle} \\
&= \frac{\langle \alpha(T_r - \mathbb{E}(T_r)) + \beta(SM_r - \mathbb{E}(SM_r)) + \varepsilon, T_r - \mathbb{E}(T_r)\rangle}{\langle T_r - \mathbb{E}(T_r), T_r - \mathbb{E}(T_r)\rangle} \\
&= \alpha \frac{\langle T_r - \mathbb{E}(T_r), T_r - \mathbb{E}(T_r)\rangle}{\langle T_r - \mathbb{E}(T_r), T_r - \mathbb{E}(T_r)\rangle} = \alpha.
\end{aligned} \tag{18}$$

and analogously $\beta^\dagger = \beta$. Therefore, doing a linear regression will lend us the correct result.

B Parameters

For the empirical experiments, we use $C_T = 10$, $C_{SM} = 1$, $C_0 = 150$, $C_1 = 0.015$.

More specifically, T_{in} and SM_{in} follow the distributions

$$T_{in} \sim \mathcal{N}(288, 100) \quad \text{and} \quad SM_{in} = 2 * s^5 \quad \text{with} \quad s \sim \mathcal{U}(0, 1). \tag{19}$$

A plot of the distributions is shown in Figure 2. We sample 100k points from these distributions and calculate T_r, SM_r and NBP following Section 3. We run a linear regression on either the full data or only the data points where the product of SM_{in} and T_{in} is above the 50^{th}, 75^{th}, 90^{th} or 95^{th} percentile.

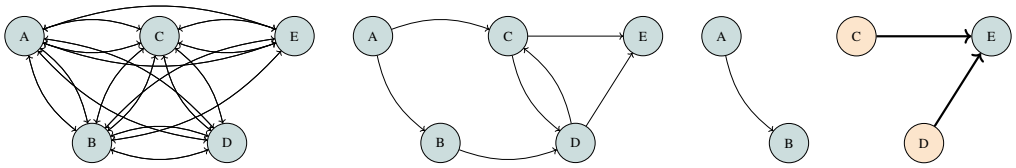

Figure 3: Example of a densly connected graph (on the left) and a simplification (in the middle) that results from ignoring some connections. On the right, we see the graph during an extreme event on C and D. Here, more processes become negligible, for example, the loop between C and D, which otherwise makes the inference of the effects very difficult.

C A Causal Perspective

In Earth's system science, causal graphs are a useful tool for understanding complex relationships between variables [5]. A causal graph is a graph in which variables are represented as nodes, and directed edges between nodes indicate the direction of causal influence. In other words, an edge $A \to B$ indicates that intervening on A has a "direct" effect on B, meaning that if we change A while keeping all variables except A and B constant, B will also change. In these graphs, many calculations are infeasible. To be able to make inference, we have to decide which connections are significant and which connections can be neglected. The graph might, however, remain complex after removing insignificant links and calculations might yield results that are strongly influenced by noise. The threshold of what is negligible depends on the strength of an effect in addition to the level of noise. During extreme events, we normally find that the influence of one (or a group of) variables is very strong, which raises the threshold as the relative influence of other processes shrinks. This leads to an even sparser graph, where some of the processes might be easier to extract.