EGUsphere, community comment CC2
https://doi.org/10.5194/egusphere-2022-541-CC2, 2022

Reply on RC2

Antoine Chambaz

Community comment on "Forecasting the cost of drought events in France by super learning" by Geoffrey Ecoto and Antoine Chambaz, EGUsphere, https://doi.org/10.5194/egusphere-2022-541-CC2, 2022

We thank the reviewer for their report.

- **In my opinion, this manuscript is not suitable for publication in its current form, particularly in a journal focused on natural-hazards research. My primary concerns can be summarized as follows:**
- **1. The manuscript is missing many key details, such that it is not possible for me to determine whether the methodological approach adopted is sound. For example:**
- **1a1. The data section does not describe the data used from different data sources and how these data compare in terms of resolution, quality etc.**

The "Data" section spans approximately three pages. It consists of 77 lines of text and one figure. We disagree with the statement "[t]he data section does not describe the data used from different data sources". However, the reviewer makes useful suggestions to enhance significantly the description. Specifically:

- **1a2. The data section does not describe [...] how these data compare in terms of resolution, quality etc.**

We thank the reviewer for pointing out the need for clarification. In Section 2.1 ("Data provided by CCR's cedents"), we now explain that the insured goods and claims data gathered by CCR over the years are **accurately** geolocalized. We also detail the resolution of each covariate when that is relevant. As for the covariates' quality, we obtained them from the National Institute for Statistical and Economic Studies (INSEE), Geographic National Institute (IGN), French Geological Survey (BRGM) and Météo-France, four trusted public organizations that notably collect, share and analyze information about the French economy and people (INSEE), geography (IGN), geology (BRGM) and meteorology (Météo-France).

- **1a3. There is no attempt made to justify the variables included in the algorithm or their significance, leaving the reader uncertain as to whether exclusion or inclusion of more variables would have improved the performance of the algorithm. Arbitrary divisions of the data (e.g., proportions of buildings built are categorized in different time intervals) are not explained and supported.**

- In Subsection "About the city-level costs of drought events" we explain our need to derive provisional city-specific costs, and how we proceed to compute them.

- In Subsection "About the city-level description" we now clarify the source of each covariate. The climatic zone and seismic zone covariates are defined by law, and we now point to the relevant articles. The proportions of buildings are defined by, and obtained from, INSEE.

Overall, as explained at the beginning of the subsection, "[a] city's multi-faceted description attempts to capture all the city's traits that, beyond the city-level SWIs (...), can explain the cost of a possible drought event." We did our best to include covariates with a potential for being relevant for the task at hand. Of course we cannot guarantee that all of them are useful.

Interestingly, some of the base learners included in the library of algorithms upon which our discrete and continuous overarching Super Learners rely incorporate data-driven routines to select more relevant covariates. Moreover, as explained in Subsection 4.1 ("Their library of algorithms"), "some of these base learners are combined (upstream) with screening algorithms. A screening algorithm is merely an algorithm that selects a subset of the covariates deemed relevant to feed the base learners. (...) In our study, we only use deterministic screening algorithms based on expert knowledge."

In Section 5 ("Discussion"), we acknowledge that the quality of our predictions strongly depends on the quality of the local description of the drought event. We discuss how the local descriptions could be improved in future work.

- **1b. There is no attempt made to introduce the very basic high-level concepts of super learning to unfamiliar readers, even to explain the concept somewhat succinctly in the abstract. This is not appropriate, given that the targeted journal is focused on natural-hazard research. Furthermore, it seems to me (based on the description provided in Section 4.2) that the authors may be evaluating the performance of the algorithm based on training (rather than test) data, which would not be appropriate.**

- super learning

The guidelines about the abstract recommend that it be "short, clear, concise". If possible, it would be good indeed to include a brief description of what super learning consists in. We give it a shot in the revised version. The new abstract now reads:

"Drought events are the second most expensive type of natural disaster within the French legal framework called the natural disasters compensation scheme. In recent years, drought events have been remarkable in their geographical scale and intensity. We develop and apply a new methodology to forecast the cost of a drought event in France. The methodology hinges on super learning (van der Laan et al., 2007; Benkeser et al., 2018). Super learning is a general methodology to learn a feature of the law of the data identified through an ad hoc risk function by relying on a library of algorithms. The algorithms either compete (discrete super learning) or collaborate (continuous super learning), a cross-validation scheme allowing to determine the best performing algorithm or combination of algorithms, respectively. Our super learner takes into account the complex dependence structure induced in the data by the spatial and temporal nature of drought events."

The third paragraph of Section 3.1 ("Presentation and theoretical performance" of the One-Step Ahead Sequential Super Learner, OSASSL) summarizes what super learning consists in. The detailed description of how we implement our two OSASSLs complements the

summary.

Following the reviewer's suggestion, we also added the brief description of super learning from the abstract in the first paragraph of Section 3.1.

- on training OSASSL

Super learning hinges on cross-validation to evaluate and compare the risks of the various algorithms. In the simpler case where one learns from independent and identically distributed data, one often implements V-fold cross-validation: first, the data set is split into V groups of roughly equal sizes (the "folds"); second, every algorithm is trained and tested V times, once for each fold, which is used for testing after the algorithm has been trained using all the other folds; third, the cross-validated (empirical) risk of the algorithm is defined as the average of the V fold-specific (empirical) risks obtained by testing.

In this study, we learn from a (short) time-series (with time-specific observations consisting of many dependent data-structures). We thus cannot rely on V-fold cross-validation. Instead, we rely on a sequential cross-validation scheme: sequentially at each time t, for each algorithm: all data till time (t-1) are used for training and the t-specific data are used for testing; the t-specific cross-validated (empirical) cumulative risk of the algorithm is defined as the average of the tau-specific (empirical) risks (where tau ranges between 1 and t) obtained by testing.

If the reviewer thought that could be useful for future readers, we would gladly include the two above paragraphs in the manuscript.

Furthermore, as explained in Section 3.1 ("Presentation and theoretical performance"), the theoretical analysis of OSASSL carried out in a companion study reveals that OSASSL manages to make up for the shortness of the time-series thanks to the manyness of each time-specific observation provided that the latter are only slightly dependent.

- **1c. There is no justification provided for the authors' exclusive focus on cities. Why not also include the costs of droughts in rural areas, when total drought costs are available (according to Figure 1)?**

We thank the reviewer for pointing out the fact the word "city" may be misleading. We certainly must clarify that all French "communes" are considered.

According to the Cambridge Dictionary, the word "commune" can be translated to "town" or "village". However, the first definition of "town" in the same dictionary reads "a place where people live and work, containing many houses, shops, places of work, places of entertainment, etc., and usually larger than a village but smaller than a city". We finally opted for the use of the word "city" regardless of the location and size.

Would a note on the choice of the word "city" and on the fact that all "communes" are considered solve the issue?

- **1d. How is inflation factored into the observed costs, particularly those from many years ago? How can future changes in exposure and population be integrated into future projections of drought costs from these algorithms? These questions should be answered clearly within the text.**

We use "constant euros". This has been clarified at the very beginning of the manuscript.

In Section 3.1 ("Presentation and theoretical performance") we make a stationarity assumption on the mean conditional cost given the (a,t)-specific collection X_{a,t} of

covariates describing city a on year t and the city-level SWI Z_{a,t} describing the drought event that year. In words, we assume that the mechanism that produces a cost after a drought event conditionally on (X_{a,t} , Z_{a,t}) does not depend on (a, t), that is, remains constant throughout time and France. In view of the reviewer's question, we emphasize that (X_{a,t} , Z_{a,t}) includes (and is not limited to) a measure of exposure and a description of the population.

Under this stationarity assumption, we can use the estimator of the mean conditional cost to make predictions at any (x,z) provided that (x,z) falls in the domain of the observed (X_{a,t} , Z_{a,t}). Of course, the closer (x,z) is to the border of that domain, the less reliable is the prediction. Moreover, if (x,z) falls outside the domain then, although a prediction may be made nonetheless, it cannot be trusted. So, in view of the reviewer's question and of climate change, not-too-distant-future projections of drought events can be made.

- **2. If the problem being tackled is "less challenging" than that of a previous study (as implied by the authors in line 270), then I am doubtful on what (if anything) the present study is contributing to the state of the art in this field.**

Forecasting the cost of drought events in France is an important task for CCR. For a given year the task will be carried out several times because, as time goes by, more relevant information is available.

At first, it is necessary to predict which cities will make a request for the government declaration of natural disaster for a drought event. Later on it is known that some cities did make the request and it is still necessary to predict for the others. Later still it is known exactly which cities did make the request. Note that once a request is made, there is no uncertainty **for CCR** about whether or not the city will obtain the government declaration of natural disaster for a drought event.

Therefore CCR currently addresses two sub-problems separately: sub-problem 1 consists in predicting which cities will make a request for the government declaration of natural disaster for a drought event; sub-problem 2 consists in predicting the cost of a drought event for those cities that obtained the government declaration of natural disaster for a drought event. In this study, we focus on sub-problem 2. On the contrary, Charpentier et al (2021) address the two sub-problems as one single problem.

Our algorithms are useful early on, when it is still necessary to predict which cities will make a request for the government declaration of natural disaster for a drought event. In that case, another algorithm (a solution to sub-problem 1) is used to predict which cities will make a request and the prediction of costs is carried out for them. Our algorithms are also useful later on, when it is known exactly which cities did make the request. In that case, of course, the other algorithm is not useful.

- **3. As seen in Figure 1, the claims data does not adequately represent the full cost of the droughts in any given year. If the purpose of the algorithm is to predict claims data, then this might be acceptable but if the purpose of the algorithm is to predict overall drought costs, then these do not seem reasonable training data to me.**

We do want to predict overall drought costs. Moreover, even if we aim to forecast the cost of drought events from year t on year (t+1), the cost of the damages in a city caused by a drought event that happened on year t is still unknown on year (t+1). In Section 2.3 ("City-level data processing", subsection "About the city-level costs of drought events") we explain how city-specific costs are estimated in such a way that the sum of all the city-

specific costs equals the overall cost estimated by actuarial studies.

- **4. I am generally concerned by the arbitrary equivalence of droughts and natural disasters. Droughts are not the only natural disasters that France suffers, yet this seems to be incorrectly implied in a number of cases:**

We should have clarified that we focus solely on drought events.

- **4a. Inputs to the algorithm include indicators on whether there have been (successful) requests for government declarations of natural disasters -- these declarations do not necessarily indicate the occurrence of a drought.**

We now use systematically the expressions "make a request for/obtain the government declaration of natural disaster **for a drought event**".

- **4b. Figure 5 shows errors for regions where natural disasters (rather than specifically droughts) occurred.**

See the two above replies.

- **5 More minor (but still important) concerns:**
- **5a. I cannot find a precise description of the aim of the study in the Introduction. (This is implicit but should be explicit for clarity).**

Thank you for noting this. We clarified of objective in the introduction.

- **5b. Figure 3: The real costs shown in this figure do not seem to align with those shown in Figure 1 (e.g., the 2017 cost of >900 million shown in Figure 3 is not found in Figure 1). So what real costs are being shown here?**

The real costs are reevaluated every quarter. We will make sure that we use the latest real costs in both figures.