Review of Levavasseur et al. – The Cryosphere Discussion 4: 2233-2275

The discussion article attempts to quantify the performance of different empirical models to predict / hindcast permafrost distribution based on climate model outputs. I found several major issues in the discussion article: (1) The manuscript is hard to read because of often imprecise wording and confusing article structure. (2) The quality of the input data is a major issue. (3) Some of the modeling decisions are inconsistent or hard to follow.

Regarding (2), the authors need to demonstrate that the permafrost maps used for model calibration and performance assessment are of good enough quality for the purposes of this study. I am not convinced that the authors will be able to demonstrate this in a revised paper.

Regarding (3), some important modeling decisions are incomprehensible to me. For example, why don't the authors use a multinomial logistic GAM? It appears arbitrary that the GAM is used to predict temperatures, and the MLR to predict probabilities. In the current paper it is impossible to attribute differences in model results to either the nonlinearity of the GAM or the choice of modeling temperatures versus modeling permafrost classes; this is certainly undesirable and unnecessary. Another major issue is the inclusion of ETOPO2 elevation as a predictor variable in models that also include SAT from climate models. Since SAT is already included in the models, the only purpose of elevation can be to represent residual altitudinal effects that are neglected by the climate models' underlying coarse topography. Consequently, only the residual elevation (climate model elevation minus ETOPO2 elevation) should be used as a predictor variable for local bias correction. I disagree with the use of elevation per se. In addition, if I correctly understood the cross-validation approach proposed by the authors, I object to its use in its present form, and urge the authors to perform a cross-validation in the statistical sense of the word.

Detailed explanations follow.

Abstract: The focus of this paper is on quantification – the reader should therefore have an opportunity to find some "hard", dependable numerical results in the Abstract: the overall measures of model performance (maybe the kappa coefficients) for each of the key models and both time periods.

p. 2235, l. 20 – "the permafrost index" – what permafrost index? Should probably read "a", and a characterization of the permafrost index should be provided

p. 2234, l. 22 – "We start with the hypothesis from Renssen and Vandenberghe (2003) that permafrost depends solely on surface air temperature" – (1) This doesn't seem to be stated as a hypothesis in the cited paper. (2) If it was a hypothesis, there is enough scientific evidence that it is wrong (e.g., numerous papers studying mountain permafrost distribution as related to MAAT, PISR, vegetation). (3) The authors probably want to state this as an *assumption*, not hypothesis (not only here but throughout the text). On the other hand, what is "surface air temperature"? MAAT, the entire temperature time series, summer temperature, daily maxima, ...?

p. 2236, l. 4-5 – "downscaling methods, bringing local information" – unclear if this refers to additional (non-SAT, maybe albedo or vegetation etc.) information, or simply to local ground-truth information that is used for local calibration? It seems to refer to the former (higher-resolution elevation data).

p. 2236, l. 15 – “smart” – omit this word, it gives the sentence the flavour of expressing a preconceived opinion on the predictive performance (and the “intelligence” of the internal functioning) of the proposed methods.

p. 2237, l. 8-10: Sentence “[The] GAM ... (a discrete variable).” – Meaning of this sentence remains unclear to me. The GAM is not limited to continuous predictor and response variables. Logistic GAMs can deal with binary response variables; any GAM can incorporate categorical predictor variables by representing these as dummy variables. GAMs are simply nonlinear extensions of GLMs, and the logistic GAM is a nonlinear extension of logistic regression. Valid arguments in favour or against the use of GAM vs. GLM would be related to the assumed or known existence of nonlinearities in empirical relationships of physical processes (or knowledge of the lack of such nonlinearities), or the greater tendency of more flexible models to overfit to the training data, if prediction is the main goal. A recent publication of Brenning (2009, in Remote Sensing of Environment) in the context of periglacial landforms suggests that, at least in this particular context, GAMs tended to be slightly better than GLMs, although they were not necessarily better than other linear methods. See also Luoto and Hjort (2005, in Geomorphology).

p. 2237, l. 12-13: “In climatology... wet or dry day sequences ... or vegetation types” – this is quite different from permafrost presence / absence modeling. For example, Lewkowicz & Ednie (2004, in Permafrost and Periglacial Processes) used logistic regression to map mountain permafrost – this example would be more relevant to the present study than the cited examples and references.

p. 2237, l. 17 – “hypothesis” → “assumption”; I do not agree that steady-state climate and equilibrium conditions are or have to be assumed when using empirical models. This would more likely be the case when using (simplified) physically-based approaches, or when attributing physical interpretations to empirical findings.

p. 2238, l. 1 – omit this sentence

p. 2238, l. 4 – “freezing point of water” – permafrost is usually defined to be at or below 0 degrees Celsius for a certain time period, which is not exactly the same as the freezing point of water, which depends on pressure, salt content, etc. See e.g. p. 83 of the cited reference (French, 2007).

p. 2238, l. 5 – 8: Insufficient detail on the origin, type and quality of “present” permafrost distribution data is provided. In addition to the original scale (or resolution) at which the map was prepared, it is essential to explain what method was used by its authors. Also, what are the uncertainties, possible or known systematic biases, regional differences in the product’s quality?

p. 2238, l. 10-11: Again, more details on quality and origin of the reconstructed permafrost extent are needed; any relevant information from the cited papers should be repeated here because it is essential for judging the utility of the data for the present article. Stating that both maps “have been drawn [sic!] in a similar way” is not sufficient and does not help support the implicit claim that these maps are of sufficient quality, which needs to be demonstrated.

p. 2239, l. 2 –GAMs are not limited to continuous response variables. Logistic GAMs for binary responses are readily available (for example, in the R packages gam and mgcv, probably the most widely used GAM implementations), and approaches for converting binary into multi-class classifiers are well documented in the statistical literature.

p. 2239, l. 19 (and elsewhere) – “temperature” – should read “MAAT” (mean annual air temperature)

p. 2239, l. 19-21 – why using this particular relationship? What is it based on? Is there any reason to believe that it is more likely transferable to the LGM than the other decision rules that have been proposed? LGM permafrost predictions will critically depend on these thresholds. Given the large uncertainties inherent in the thresholds, it would appear reasonable to perform a sensitivity analysis or compare the results obtained with different decision rules.

p. 2240, l. 1-5: This is not convincing. (1) Other variables (such as PISR, vegetation) are not empirically compared to permafrost distribution, and consequently the implicit claim that they are not important cannot be made. (2) The claim of an “obvious” empirical association between the permafrost map categories and MAAT should be supported by numerical evidence, i.e. suitable measures of association (AUROC, kappa and the like). On the other hand, it may be acceptable to simply state that MAAT is the only variable considered, that other influences are known to exist (provide references), but act mainly at the sub-pixel scale, and that the model’s predictive performance will tell us in an objective way if the results are of an acceptable quality.

p. 2240, l. 10: “is a reasonable approximation” – Such an affirmation is not acceptable in this part of the paper because the question whether and how reasonable this approximation is constitutes the main goal of this article.

p. 2240, l. 20: “cubic splines” – in the general formulation of the GAM, the nonlinear transformation functions are not necessarily cubic splines. Use a more general expression here, and specify later the particular settings (cubic splines) used in this study.

p. 2240, l. 25, “gaussian" → “Gaussian”

p. 2241, l. 8-10: Cross-validation is a well-defined and widely used statistical estimation technique that is based on resampling (partitioning) the observations, not the variables. It is incomprehensible to me what the described leave-one-variable-out cross-validation would do to help assess the technique’s predictive performance. And again, to be clear, this is not a cross-validation.

p. 2241, l. 14: please see the citation() function in R – it provides recommendations for referencing R and its packages.

Section 3.2.1 in general: provide a rationale for the choices you make, instead of simply stating that only one physical predictor is used and that you “choose to work with nine”CMs, for example.

p. 2242, l. 1-2: SAT is mean annual SAT?

p. 2242, l. 2-4: “This variable ... bilinearly interpolated at 10’ resolution” – This tells me that the actual downscaling takes place here, by interpolating SAT. The GAM / MLR simply use this finer-resolution data to calibrate an empirical relationship, but they themselves do not perform down scaling.

p. 2242, l. 13-15: Not present-day ETOPO2 elevation (or LGM elevation, respectively) should be considered here, but the difference between ETOPO2 and the elevation used by each CM should be used in order to reflect elevation-related temperature biases.

p. 2243, l. 15-24: This belongs into the Methods section. The decision to mask certain areas is part of the chosen method, not a result of the application of the chosen method.

p. 2244, first paragraph: I strongly recommend starting with a general quantitative performance summary before presenting the maps and analyzing regional / altitudinal difficulties of the model.

p. 2244, second paragraph, to p. 2247, l. 2, and again p. 2247, l. 24 to p. 2248, l. 27: Introduce the performance measures and the MLR in the Methods section, not here. This will make the results section much easier to read.

p. 2244, l. 16: "To quantitatively assess the effect of our downscaling on permafrost representation, we measure the agreement between CMs and data..." – This sounds odd to me. First, what is "data", in this context. Second, if you want to assess the performance of the downscaling methods (=GAM, MLR), why do you say in the same sentence that you are going to assess the CMs, and not the downscaling methods?

p. 2244, l. 24: "CMs underestimate the permafrost area" – well, CMs do not produce permafrost maps, but SAT maps, do how can they underestimate the permafrost area? Are the temperature thresholds of Renssen and Vandenberghe (2003) being used?

p. 2244, l. 26, and elsewhere: more generally used terminology is available to refer to %CP and %DP; e.g., concepts such as true positive rate (or sensitivity), or, using remote-sensing terminology user and producer accuracy. Using such general concepts would make the results much more readable.

p. 2245, l. 2: "...show the limits of the GAM method" – clearly, this is a discussion item, should not be presented as part of the results. I am also not convinced that this is an issue of the GAM, but maybe of the authors' choice of applying the GAM to temperature prediction and then applying fixed decision thresholds.

p. 2245, l. 9 – I do not understand this interpretation.

p. 2245, l. 10 – standard deviation based on 9 observations, one of which (in the LGM situation) is pretty far off the mean value? I am avoiding the word "outlier" here, would rather interpret this as either supportive of a skewed distribution, or simply a small-sample effect. In any case these standard deviations are not very reliable statistically, they will be sensitive to the deviation from mean in any individual observation.

p. 2246, l. 9 – 13: motivation and calculation of kappa_max remains unclear to me.

p. 2247, l. 2 – "statistical significance" – use this expression only in the context of statistical hypothesis testing.

p. 2247, l. 4 – "statistically relevant, in better agreement with data and not by chance" – The authors should distinguish between two questions: (1) is the GAM (or any other method) better than chance agreement achieved by, e.g., tossing a coin? (2) Is the GAM better than any of the other methods studied? The kappa coefficient is one possible approach for answering question (1) as it (is intended to) adjust for chance agreement. In the case of pairwise comparisons between different method (e.g., GAM

versus MLR), it is harder to tell whether the estimated differences (in kappa coefficients, for example) can be attributed to random variations, i.e. chance. In my view, in this context most readers would interpret “not by chance” in the sense of statistically significant (sensu hypothesis testing), although even this would be problematic and not straightforward. In brief, the authors should be aware of the complex statistical implications of the apparently simple statement provided on l. 4, and should therefore refine the wording not only here but throughout the text in order to make sure that all interpretations are unambiguously supported by the results.

p. 2248, l. 19: why reference a GAM paper (Yee & Wild, 1996) and R package for logistic regression? Was the VGAM package used for MLR in this study? If yes, say so.

p. 2249, l. 27: please present quantitative results that support this statement.

(Note: My comments for the following manuscript pages are somewhat reduced because most of the previously made comments also apply to the LGM-related section.)

p. 2252, last line, to second line of p. 2253: application of Fisher’s test and Student’s test to “variances from MLR and GAM-RV simulations” – unclear what “simulations” (or predictions) it refers to, and how these are tested. It has to be made clear what is tested for what reason, and if the test’s distributional assumptions are honoured. The latter is unlikely here because the observations used for model fitting have to be assumed to be spatially autocorrelated.

p. 2253, l. 2-3: Even if the previously mentioned test results are statistically valid, the statistical non-significance does not show anything. Only positive (i.e. significant) test results should be interpreted, the inability to reject the null hypothesis shouldn’t. However, even if, I do not see how this would allow an interpretation in terms of [reconstructed or hindcasted?] permafrost distribution being “strongly driven by the large-scale temperatures from CMs.”

p. 2253-2254 (itemized list): should be worked into a new Discussion section. In the Discussion, the present results and methods should also be discussed in the context of the broader literature.

Fig. 5 – The box-and-whisker plots appear to be based on nine values each? I am not convinced that a box-and-whisker representation should be applied to such a small sample. I recommend omitting the box-and-whisker plots, and showing only the point symbols and for each model a line representing the median value.

Fig. 6e) – The modeled relationship between discontinuous permafrost probability and annual mean temperature is clearly nonlinear and non-monotonic – how can this be accomplished by the MLR, which can only produce monotic relationships?

Wording:

Some language editing is required (grammar, diction), the following list is not complete.

Throughout the text: “inter-variability” → “differences” [generic term], “inter-variation” [rarely used], “co-variation”, “inter-relation” [may have a stronger interpretative or even causal connotation, which

should probably be avoided in this context]; “variability” refers to the “ability” to vary, i.e. (in a statistical/stochastic contgext) to a property of the underlying random variable, not to empirical findings of very limited coverage.

p. 2234, l. 19 – “proves” – avoid this strong word, I doubt that the comparison of several semi-physical models can “prove” the stated relationship between two physical phenomena, permafrost and SAT. Also, is it not trivially true that the permafrost – SAT relationship is not simple? Replace this statement with one that expresses a positive finding, e.g. that a particular relationship or process is particularly well represented by the chosen model(s)

p. 2234, l. 24: “Our SDMs do not significantly improve permafrost distribution” – of course they don’t, they are models, not nature. Change wording: “...the prediction of... compared to...”

p. 2234, l. 25: “at this period” – omit, and write “LGM” earlier in this sentence.

p. 2235, l. 11 – “local” – refers to sub-pixel / sub-grid-cell (I think so), or to locally varying as opposed to global (e.g., local trends across several grid cells, depending on particular geological or land-use conditions etc.); in l. 17, “local” seems to refer to “spatially varying”; on p. 2236 l. 14-15, “local” appears to mean “sub-grid-cell”

p. 2236, l. 1 – “resolved” should presumably read “solved”

p. 2236, l. 1 – “need a lot of computing time” → “are computationally intensive”

p. 2237, l. 7 – “simulated” should read “predicted” (same for most other occurrences of simulate / simulated elsewhere in the text, e.g., p. 2248 l. 6)

p. 2237, l. 19 – “simulating” → “hindcasting”

p. 2237, l. 21 – “their” – whose? The paleo-environments of the CMs?

p. 2237, l. 23 – here and elsewhere in the text, the word “data” should be replaced with more precise, context-specific words. Here, for example, it may refer to direct or indirect observations of the climate variables to be predicted, so “observations” or “observational data” might be a better word choice; “data” could be anything.

p. 2238, l. 5 – “geological” → “cryospheric”, or simply omit the word; not all permafrost researchers are geologists or consider their research or methods to be geological.

p. 2239, l. 2 – “simulate” → “predict”

p. 2244, l. 23: “decrease of permafrost area” – use more precise wording; maybe the “smaller predicted permafrost area” or “negative bias in the predicted total permafrost area”; similar issues with “decrease” and “increase” elsewhere in the text

throughout the text: “our” → “the”

p. 2252, l. 7-8: “GAM slightly warms temperatures” – sloppy wording, the GAMs are not to blame for rising temperatures.