Dear Dr. Currie,

Thank you for your consideration of our manuscript, “Cross-scale integration of knowledge for predicting species ranges: a metamodeling framework.” We appreciate the work of both editors and the reviewers, and we believe their comments have greatly improved the manuscript. We are pleased to submit a revised version addressing the comments of the reviewers and editors. For convenience, we have indicated revised sections of the main manuscript in a blue font. Additionally, we provide below the text of reviewer/editor comments (in italics) followed by our response and a summary of the changes to the manuscript (in upright text).

Thank you again for your consideration of our manuscript. We look forward to your response.

Sincerely,

Matthew Talluto

Laboratoire d'Ecologie Alpine, UMR-CNRS 5553

Université Joseph Fourier

BP53, 38041 Grenoble cedex 9, France

**EIC Comments**

*This paper was submitted in the Concept Paper category.  There is also a category called Macroecological Methods. It seems to me that this is a fundamentally a methods paper.  Unless you have strong objection, I will move it to that category. The only operational difference is the heading under which the paper would eventually be published.*

We have no objections, and we have changed the category with the submission.

*Table 1 -- Tables and figures should be understandable without reference to the main text. This legend is rather cryptic.  The table contains several undefined acronyms.*

We have updated the table legend following these guidelines.

*Fig. 1 -- Is the reader supposed to infer that precipitation below 0.1 has not been sampled? The points in the panel on the left extend beyond the range of the axis.   Is there a reason that you have not extended the axis?*

We have extended the axis to 0 on the left panel, and clarified in the legend that precipitation was only sampled between 0.1 and 1.

*The publisher charges to print colour figures, although there are rather few print copies any longer.  Fig. 3 can be used in colour in the the digital versions of the paper, and printed in black and white, at no cost. Information would be lost from Figs. 5 and 6 if we did this, because these scales are dark in the middle and light at both ends, or vice-versa.  If you can produce a figure with a colour scale similar to the one in FIg. 3, or a black and white version of this figure, then we can proceed as above.  I am unsure about Fig. 4.  Another option would be to pay the colour publication fees.*

We would like the figures to appear in colour, and will pay the fees.

**Handling editor comments**

*\* Global Ecology and Biogeography uses structured Abstracts; please adapt your current text to the journal style.*

We have revised the abstract as requested

**Reviewer 1 comments**

*•Scaling results from submodels to inform the metamodel seems to be a crucial part of the framework, and also one of the most delicate (L135-151, and L338-351). I am worried that fitted relationships at one scale (e.g. the ecological experiment relating growth rate to precipitation in example 1) may differ substantially from relationships at larger scales (i.e. those relating climate to species distributions). In other words, observed relationships between climate and species performance or distribution may be  
scale and context-dependent (e.g. McGill 2010). Hence, I am concerned if constraining large-scale models on results from small scale experiments might be ‘dangerous’ and actually provide poorer predictions.  
  
To illustrate with their example 1, I wonder if the relationship between population growth rate and precipitation as fitted to the experimental data might not hold when scaled up to the range scale. I can foresee many situations in which this could happen: for instance, if response to precipitation varies depending on temperature or the environmental context (e.g. if temperatures are higher, much higher precipitation would be required to achieve the same growth rates). Thus, even though the authors put great care on downweighting results from this submodel to the metamodel scale in their example (L161-164), I am concerned that the crucial step of extrapolating or  
transferring information from local-scale, experimental data, up to range scales may not be so straightforward. It would be very useful if the authors could provide more clear guidelines and thoughts on how to ensure this scaling is made properly.*

The reviewer raises an excellent point (and indeed, one that is valid whenever attempting to model large-scale ecological patterns as a function of smaller-scale processes). The Discussion touched on these issues briefly in the original manuscript. We have significantly expanded this section (**lines 379–390**) and emphasized that the users of such models should take care in implementing scaled models, and err on the side of caution by downweighting models when there is uncertainty about the appropriateness of the scaling used. We have also added a section in the Appendix providing some practical guidance on implementing model weights.

*•Related to this, given that weighting and prior definitions are a crucial step in this workflow (L161-164), it would be good if the authors provided a slightly more detailed explanation of the weighing procedure, either in the main text or in the appendix. Of course the details could be searched in the R code, but I think a textual explanation would be good too.*

A description of the weighting procedure, including guidelines on selecting and applying the model weights, now appears in the “Implementation” section of Appendix S1.

*•The description of the methodology for example 2 is overall much less clear than example 1 (considering the appendix too), and should be improved by giving more details and more clear explanations of the procedure. Particularly unclear to me are the future projections (“Integrated-Future”) (L214-217). It is unclear if Phenofit projections are used to constrain the naive model, or the results of the naive model  
are used as priors for a metamodel based on simulated datasets arising from Phenofit projections (page 37, L44-45). How are Phenofit projections and the naive model integrated for the future, when obviously there are no occurrence data to fit a naive model? In addition to improved textual explanations, more explicit details about the statistical structure of the models need to be given, as was done for example 1.*

Thank you for pointing out this oversight. In the interests of remaining under the maximum word count, we have added a brief additional explanation in the main text (**lines 221–231**), as well as several additional paragraphs (and a more complete statistical treatment) in Appendix S1. The briefly summarize, the naïve model was a static GLM relating present climate to occurrences. The Integrated-Present model combined this model with the present predictions from Phenofit (with pseudo-occurrence data generated via simulation), fit to present climate. The Integrated-Future model did the same thing, but used future predictions from Phenofit fit to future climate. The confusion on this point is also related to the confusion on the next point, regarding the response curves; we hope the revisions on this point and on the following help clear up any confusion on both points.

*Also, I do not understand why response curves (Fig. 4) should change between the present and the future. Aren’t they supposed to represent estimates of the climatic niche of the species, considered static as evolutionary processes are not considered here, as far as I know? Aren’t predictions of future distributions just based on fitted relationships between present distribution and climate (i.e. fitted for the present, projected on the future)?*

We have added some text in at the end of the **Example 2** section clarifying these points (**lines 251–265**). The response curves are indeed estimates of the realized niche with respect to two macroclimatic variables. There are two reasons why these may change between the present and future models. The first is that these are simply statistical estimates of the niche, and incorporating different datasets (e.g., present vs. future predictions from Phenofit) will change those estimates. The second explanation is that the realized niche with respect to these variables may well change in the future if the smaller-scale processes responsible for limiting species distribution do not respond linearly to macroclimate. It is exactly these small-scale processes that Phenofit attempts to model, making its future predictions an appealing information source for model integration.

*Typos:  
• L183: Will scripts be Appendix S2 rather than S1?  
• L194: Appendix S1 states 7 climatic variables initially. Please clarify.  
• L214: 2100 or 2080? See L186, and Appendix S1.*

These have been revised; L183 now correctly refers to Appendix S2, L194 refers to 7 variables, and all references to 2080 have been changed to 2100.

**Reviewer 2 Comments**

*(1)     The presented approach is not the first and only application of hierarchical Bayesian modelling approaches in species distribution modelling integrating correlative and mechanistic models (cf. e.g., dynamic range models presented by Pagel and Schurr 2012). Therefore, the approach needs to be presented in relation to this approach (and maybe even additional) approaches.*

We have added a section in the Discussion (**lines 274–284**) addressing how our framework compares to Pagel and Schurr (and other hybrid approaches). To summarize, Pagel and Schurr linked a demographic model with the observation processes, as well as occurrence and abundance data. The approach is powerful and conceptually quite similar to ours, but requires a fully parameterized demographic model. We integrate models at the scale of the predictions, thus fully parameterizing a detailed process-based model is unnecessary; the user is free to use what data are available.

*(2)     Compared to classical statistical modelling approaches such as GLMs, hierarchical Bayesian models are a comparatively complex method. This becomes obvious in the appendix as well as in the quite elaborated R-scripts provided. For being accessible for a larger set of users including conservation biologists it might be a good idea to add an even simpler introduction into the underlying principles – preferably by an additional appendix or at least by referring to some introductory text books. Appendix S1 is a nice introduction into the presented approach, but for those not familiar with this methodology, it seems to be still too complex for the non-experienced. The extremely valuable approach will certainly find more users, if the manuscript provides a bit more comprehensive tutorial- or working-guide-similar explanation in an appendix (see e.g. Elith et al. 2008, Elith et al. 2010). Since the R-code is already provided, it's just a small step to include a more detailed tutorial dealing with the code (but see comment no. 3 below), which would improve the accessibility of the approach enormously.*

This is an excellent suggestion. We have added an **Implementation** section to Appendix S1, detailing the steps required to run the models (see also the response to point 3 below). We have included the practical details of how to run each script and what the expected outputs are, and also directed beginners to a few general resources.

*(3)     I was not able to run the second example; as far as I understand, there's  code missing which estimates the "naivePosterior" which is used in 3\_prepare\_integrated\_data.r und 5\_process\_posterior.r (probably some code named 2\_..., and maybe also some code named 4\_...). The provided code should be complete. It would strongly benefit from some tutorial-like treatment in an additional appendix.*

Following this comment, we have thoroughly reviewed our code and found it to be complete. This appears to be the result of some confusion regarding the procedure necessary to build and run the models; steps 2 and 4 are documented in the makefile and require compiling and running a standalone C++ program, and produce the files that are needed for steps 3 and 5. To clear up this confusion, we have thoroughly documented the build procedure in the Implementation section in Appendix S1. We have also included the files produced by steps 2 and 4 within Appendix S2. This will allow users not familiar with C++ and those wishing to skip the lengthy posterior estimation to reproduce the R portions of the analysis.

*(4)     For model evaluation, the analysis focuses on AUC, i.e. model discrimination, only. But AUC is known to be less sensitive and not directly related to model calibration (e.g. Harrell 2001). Therefore, the analysis should be supplemented by additionally considering model calibration criteria such as explained deviance as well as bias and slope of the calibration curve. The AUC-values achieved in example 2 (0.72/0.73) are assessed as "relatively well" – but in fact, models with AUC-values of about 0.7 are far away from being well performing, but just sufficient. Therefore, it would be really helpful to see these additional performance criteria such as explained deviance. This surprisingly hardly satisfactory performance should also be discussed.  
What about model performance in example 1?*

We reviewed in detail our model evaluation procedures. The relatively low AUC values were in fact a result of a bug in the code we used to estimate AUC. Briefly, we computed “mean” AUC values for each model by computing the AUC for a model consisting of the posterior means of all parameters. This is incorrect, as MCMC yields joint posterior distributions; there is no guarantee (and indeed, it is quite unlikely) that the most probable univariate parameter estimates, when combined, produces the most probable joint model. The correct procedure is to compute AUC for each posterior sample and report the distributions (along with medians). We have fixed this code and report updated median values for AUC (0.802 and 0.797). We have also added several sentences to the description of Example 2 in Appendix S1 clearly explaining this procedure, and we have included a figure showing the entire posterior distributions of AUC for both models. To address the reviewer’s request for an additional model evaluation metric, we have computed posterior calibration curves and provided them in Appendix S1.

Regarding model performance in Example 1, because the data were simulated, we have no independent dataset with which to evaluate the performance. Furthermore, such an exercise could only inform on the model’s ability to fit additional simulated data. The purpose of Example 1 was to illustrate the procedure in a simplified context. The best metric for evaluating the performance of Example 1 is the comparison of posterior uncertainty. This is already included for model predictions in the figures in the main manuscript. To provide an additional illustration, we have included posterior densities of all model parameters so that readers may see the effects of integration on parameters informed with additional data (i.e., the precipitation parameters) compared with parameters using only one data source (i.e., temperature parameters). We feel that more comprehensive evaluation statistics are more useful with the second example, where they can be interpreted meaningfully in terms of the ability of the model to predict the range of a real organism.