Review log for Validation Report section "Model validation outputs for use in SEP uncertainty calculations"

Reviewer name: Michael Dietze

Start date: 2021-09-30

End date: 2021-11-02

Key Questions from Indigo for reviewer

- *Does the section meet the Model Requirements as revised by Appendix H? Enumerate the requirements that are and are not met.*
- *If the section does not currently meet the Model Requirements, what changes are needed to bring it into compliance?*
- *Are the reported uncertainty values usable for calculation of uncertainty deductions according to both SEP Appendix D.1 and D.2?*
    - *In particular, are time units handled acceptably?*

1. As a whole the DayCent-CR model meets the CAR soil model validation requirements. There are a few places where the validation statistics were not sufficient over the full domain, but in these instances the Validation report is clear about these restrictions. Furthermore, these restrictions will have little practical impact because they uniformly apply to cases of particularly large soil C changes, which are going to be associated with experiments that were longer in duration than the model is going to be applied over.
2. The one potential area of concern was the demonstration of independence of the folds constructed for cross-validation. On the plus side, this team correctly made sure to not split observations from a single site over time. On the minus side, some of the sites are fairly close together spatially relative to the overall domains being considered. Due to limited sample size of available data, and the tendency of such data to be concentrated at a limited number of academic or agency research sites. At my suggestion the team constructed semivariograms for the raw SOC pools, change in SOC, and model-data residuals. In these analyses there was no clear evidence for spatial covariance. Furthermore, the magnitude of the residual error's sill was consistent with the model's RMSE, suggesting that the RMSE was not biased low due to spatial autocorrelation.
3. Yes, the uncertainty values are usable for uncertainty deductions and time units are handled correctly. In all cases assumptions were conservative. There was an extensive discussion of the time component of the model validation statistics and the approach proposed in Appendix H is conservative with respect to time.

Recommendations made to the project proponents during review, and current status of their resolution

1. I found the initial presentation of the statistical approach used to be lacking in a range of important details (how the likelihood was constructed, convergence, sample size, posterior correlations, etc). Resolution: all requested details have been provided, including the

posterior parameter samples themselves. The raw (n=220) and effective (n ~ 20) sample size of the posterior samples is definitely on the low side, but so long as these parameter samples are used as part of posterior predictive distributions (i.e. with residual errors and random effects propagated) then we decided that this would be a conservative assumption. Similarly, the Gelman-Rubin convergence statistics seemed on the high side compared to my past experience, but evidently this is normal for the DREAM algorithm

2. With regards to the question of using the posteriors from each of the five folds vs using the posteriors from refitting to the full data set, I suggested plotting how the cross-validation differs from the full refit to demonstrate that the impact of refitting is negligible and would not have changed any of the validation statistics. Resolution: figures added showing negligible changes in prediction (Figure 54) and that most parameter estimates were consistent (Figure 55), though future calibration analyses would benefit from looking more closely at a few of the more variable parameters (P2CO2_2, TEFF2, WEFF2).
3. I also asked for greater detail about how the model was initialized and what information was derived from external databases (gSSURGO) or spin up, including when the first SOC measurements were not at the start of the experiment. Resolution: Additional details were provided showing that these aspects were handled properly
4. I asked for greater detail about the parameters that were calibrated, how they were selected, and how the global sensitivity analysis was performed. Resolution: Revisions provided details were sufficient and followed standard practices. A future recommendation would be to spend more time on eliciting more informative priors from experts, especially as around half of calibrated parameter are currently pushing up against the edge of their uniform priors.
5. I asked for additional information about how the folds were constructed and how independence was demonstrated. Resolution: As noted above, this information was provided and suggest that the folds were sufficiently independent.

Recommendations the reviewer will be sending forward to CAR

This Validation Report is the first one submitted under the SEP, and it is therefore a test of the Model Requirements as well as of this model. If your review finds places the Requirements are deficient (whether or not correcting them would require changes in the Validation Report), recommend changes here.

1. **Preregistration:** While I have reasons to trust that this specific analysis is approaching their analysis honestly, it occurred to me while reviewing that if a team dropped a small number of their calibration/validation sites where their model was performing poorly I'd have no way knowing this (dropping large numbers of sites runs the risk that a reviewer would catch that their literature data search had been done poorly). This possibility could be reduced if the data being used for cal/val had to be preregistered before the Validation was conducted. Obviously this isn't perfect, as a team could similarly have done preliminary analyses before the preregistration. This might be improved if the preregistration was communal and cumulative (i.e. once teams submit lists of sites that have public or literature data, all future teams would have access to that list and would need to provide a clear explanation about why any relevant sites/data were not included in their cal/val). This would be like having a "standard" dataset, but would allow that dataset to continue to grow and evolve.
2. **Bias Identification and Correction:** It was unclear to me if the existing allows (or encourages/rewards) more in-depth analyses of model bias. For example, if model bias varied systematically with temperature or latitude, and calibration didn't correct this bias, could a team

either (A) use the analysis of bias patterns to propose restrictions on model application (e.g. model should not be applied above a certain latitude) and/or (B) submit for Validation hybrid model that combines the process-based model with a statistical model that corrects model bias (and possibly also uncertainty intervals). Obviously, the statistical component would have to be reviewed carefully to ensure that any corrections are appropriate and not simply overfitting the training data (but this doesn't seem like a fundamentally different problem than the current validation requirements)

3. **Fertilized Soy:** Because it is rare to fertilize Soy with N (because it's a N-fixer), Appendix B proposed considering the soy CFG validated for inorganic N fertilization if Soy has been validated for other conditions (e.g. cropping, planting, harvesting) and inorganic N has been validated for another annual CFG. I would point out that this argument would be stronger if the other CFG was not a grass.
4. **Organic amendments:** Similarly, Appendix D proposes to validate this management operation holistically, rather than by CFG, under the argument that large carbon inputs from the amendments dominates the across-CFG differences in litter inputs. While this is true from an inputs perspective, and while I also appreciate that there may be sample-size arguments for lumping, it isn't necessarily true that differences in SOC accumulation across CFGs are solely a function of input rates, as crops also differ in root enzyme production, nutrient uptake, and microbial communities. For example, papers by Colin Averill have shown a ~2x difference in soil C storage in forest ecosystems depending on mycorrhizal type and N input rates, and present a solid argument that the soil C storage differences are driven by differences in N uptake between AM and EM fungi, not by C input rates.