Review of Schneider et al 2024 Optimizing climate models

I liked this article a lot.

1. I believe that you shouldn't trust a climate model prediction that you don't understand conceptually. This is particularly necessary for climate modeling because we can't validate any of our predictions until it is too late. Accepting anything on blind faith from a black box model seems like a recipe for disaster. This requires a convergence between episteme and techne which is different from your framing around line 25. Regarding line 248, Bjorn has told me that not using deep convection is strongly motivated by a desire to understand what his model is doing rather than just because it makes the simulation better (which I think most km-scale modelers at this point believe is not necessarily true)
2. I felt that the assessment of km-scale models on p. 7 (and, to a lesser extent, high-res models on p.6) was a bit unfair. My feeling is that conventional GCMs have been optimized and tuned for decades but these higher-resolution analogues are still new and generally haven't been well tuned. I think they have a lot of room for improvement. It is hard to say at this point how much benefit they will provide, but they will certainly be better than the versions you're analyzing. One particular challenge is that they are too expensive to tune, which both supports my claim that they have a lot of potential for improvement but also your point that km-scale models may be too expensive to be practical.
3. Paragraph between line 190 and 195: my personal feeling is even stronger than your argument here – I don't think it's clear that it will ever be possible to adequately parameterize clouds from variables available on the grid scale. Necessary information may simply not be available. I can't think of how to edit your text to express this, so just adding as a comment.
4. L221: When you say "conditional averaging", I think you mean that you will break terms in the governing equations into summands satisfying one condition or another. Just averaging over one particular condition (e.g. only for updrafts) generally does not result in a statement equivalent to the original governing equations. I've had postdocs go astray this way.
5. Numerated item beginning on L220: I would add that carefully applying scaling arguments to make simplifications and being explicit about the simplifications you're making is critical for readers to understand what you're doing and to be able to assess how much trust they should have in what you do. It may turn out that some assumption you make (like the PDF for subgrid variability) turns out to be inappropriate in some edge case and having those assumptions be clearly listed will help in tracking down these issues.
6. To amplify the last comment, I believe that assumed PDF shape and in particular assumed covariances between variables will be central to the skill of the kind of model you're advocating.
7. Discussing surrounding the list of desired properties for parameterizations starting around L219: I think you're missing the possibility for covariance between variables within processes and particularly sub-grid scale interactions between processes. For example, condensation is nonlinearly stronger in portions of a grid cell with stronger updrafts, which wouldn't be captured in models where condensation is performed in microphysics rather than turbulence schemes. I also like Devine et al (2006; GRL), which points out that

interactions between convective transport and sub-grid scale spatial variations in DMS are critical for getting cloud microphysics right. All these covariances between processes are things that improved resolution fixes, but would be hard to parameterize without having a single really complicated parameterization that does everything.

8. Paragraph starting on L304: It is interesting that most modeling centers have found that decreasing dx provides better simulation skill but decreasing dz generally makes the model worse (at least without a ton of extra work). I think this is because the model is actually *more* sensitive to vertical grid and gets its skill from tuning rather than accurate discretized equations, and because it is easy to make discretization mistakes in the vertical, so it doesn't conflict with your argument. But this explanation does explain why modelers have focused on improving dx rather than dz even though the latter is more cost effective, as you point out. It may also be worth mentioning the theoretical discussion about the need to change both vertical and horizontal resolution at the same time from [Lindtzen and Fox-Rabinowitz (1989; MWR)](). It is funny how nobody actually links dx and dz when changing resolution even though we know we should.
9. A minor point, but your argument that we should choose resolution based on what we can afford rather than some theoretical panacea (L325) only works if you've formulated your parameterizations in a way that works across all resolutions. Jumping from 100 km to 3 km dx was largely motivated by the sentiment that gray-zone convection must be avoided at all costs.
10. L338: Using emergent constraints in your cost function is a great idea *if you're positive they are real constraints.*
11. P. 16: it strikes me that your "ML as an inverse problem" is very similar to climate model "autotuning", which is being pursued by a lot of groups right now. It may be worth comparing and constrasting your approach against autotuning.
12. I felt like you were glossing over the difficulty of ML as an inverse problem when you have several uncertain parameterizations you are trying to optimize but only have net atmospheric state as your input and net state change as your tuning target. At best there are probably several optimal solutions and at worst your training data is insufficient to predict appropriate behavior. A bit of discussion about why you think this problem is tractable would be appreciated.
13. L421: I really like the idea that we should use ML on detailed subprocesses rather than entire large chunks of the model. I think about this often for microphysics: we have a good sense of what controls each of these detailed processes and we can see that the whole spectrum of conditions these subprocesses will face are probably already being experienced in current climate, so I have a lot more confidence using ML on them to predict future climate. I think you could go a bit further on this point by saying that the choice of which parts of the model will be replaced by ML needs to be made using expert judgement that the process of interest will be climate-invariant and sufficiently sampled in the current climate.
14. I think observational uncertainty is a critical aspect of model optimization, but you don't mention it.
15. L481: I'm a big fan of the idea that climate models need to get a lot better because they must be used for decision support. I like your framing for how we level up these codes. I

would add that exhaustive unit testing and convergence tests are also needed to provide needed confidence in our predictions.

Proofreading:

1. L56: missing “t” at the end of “Turing test”
2. L102: “balance of TOA radiative energy fluxes must also be closed” seems like an obtuse way of saying that energy must be conserved.
3. Fig 1 caption “averaged over the same simulation length” – aren’t the AMIP cases actually averaged over the same *dates*? You’re not comparing AMIP from 1979 against ICON from 2020, are you?
4. Also for Fig 1, I found “LL, MM, HH” to be confusing terminology. You could just as easily have used titles providing the actual dx for each run. It is also unclear whether you’re comparing the ensemble-average AMIP result against a single run from IFS or ICON in those right-hand graphics. I’m surprised that km-scale models have worse precip RMSE than coarse models. This isn’t what I’ve seen… which makes me wonder whether you’re comparing ensemble-mean skill (which tends to be superior to individual models) against a particular instance.
5. L200: typo – identifying should be identified.
6. Wherever you talk about the loss function, you add “(1)” afterwards. I found this distracting because I kept thinking you were going to start an enumerated list. I think all readers will know what you’re talking about just by mentioning the loss function without referring to the equation.
7. L240: Ad citations for pioneering EDMF papers rather than just your group’s recent work here?
8. L337: “This” is an unclear antecedent
9. L370: including an example where the forecast loss function is optimized but climate isn’t would be useful. It is easy to think that if you do a good job in each timestep, you will do a good job overall because climate is just the collection of timestep-level results. The obvious counter-example is if you are biased a tiny bit in the same direction every timestep.
10.