This is a well written and well reasoned manuscript that i found novel and interesting. I particularly like looking at what models do rather than what they purport to do, and this is an example of how we should evaluate what models do to determine if they are adequate for a given purpose. From this point of view if we think that the atmosphere does, in rainy regions, show a robust behavior that a model does not follow, then this is a good reason not to use that model for things that might depend on this relationship. The next step is, of course, to see what depends on this relationship, something the manuscript also takes up.

I have a couple of major comments that the authors may want to consider, but given that the review is part of the public record, they don't need to be addressed as a matter of publication. However, some of the points about editing (particularly as they are part of the ACP standard) should be addressed.

Major comments:

It is not clear to me what controls the spread within the Palmer-Singh space and what assumptions are behind this spread being aligned along a constant ϵ. For one, ϵ could vary based on environmental conditions, this is certainly the case for Nordeng-Tiedtke which has a small and large-scale entrainment, and even for a fixed entrainment it seems that the adjustment timescale might influence the joint histograms. Given this, it seems that the direction of spread is not necessarily a measure of goodness of fit. It could be argued that IPSL is well described by the entrainment rate that passes through the center of mass of points, something that is supported by the shift with warming, but that variability is expressed by ϵ varying with environmental conditions, as claimed for MPI-M with Nodeng-Tiedtke.

I don't really agree with the premise that all models are wrong, but in precisely the way that would allow their errors to be corrected by the 'emergent constraint' or 'process oriented diagnostic' . I also would refrain from introducing/using shorthand for things like POD as this gives them an air of precision that is not warranted. I would further maintain that emergent constraints have been largely a dead end that somehow encourages the fantasy that the correct answer is to be found in the garbage heap of inadequate model output, which is used more because it is available, and less because it is demonstrably adequate for purpose. On the other hand the authors's analysis is a very nice way to think about how the world works, in which context more focus on that point, and what statements about the world different models might adequately test, would be more useful. This is actually what the manuscript does in the second part, albeit burdened by the baggage of false ideology (just to phrase things colorfully)

Some minor points, some of which are particular instances of the above are provided below.

- See the ACP typesetting rules, which revers to the IUPAC standards, i.e., roman versus italic and when (operators, name subscripts, are roman) e.g., §1.3 of the GreenBook. Also I persist in trying to encourage my colleagues not to use the word heat as a noun, and not to call enthalpy heat, and not to use capitals for specific quantities… hence ℓ for the vaporization enthalpy. The subscript 'v' is not even necessary as the fusion enthalpy does not enter into any of the arguments.
- Given the major comment above, and that the relationship does measure something, for some quantities (i.e., the CAPE estimation) maybe one need not refer to only those models with $\epsilon > 0$. In particularly if there is another process that gives an offset, then the model may well follow the ideas in the ZBP modulo an offset. On the other hand for the boundary layer humidity, it would seem that only models with $\epsilon > 0$ make sense to analyze.

- Although it is well caveated, I'm a bit reluctant to read too much into the reanalyses. Their differences were not insubstantial given the similarities of models and approaches they employ. As more an more large local-area-sounding data sets become available (GATE, FGGE, EUREC4A, OTREC, ORCESTRA and so on), I think this framework could be more usefully employed outside of model space and also outside of the idiosyncratic nature of mostly land based tropical sondes used for routine weather observations.
- To show how little attention I paid to type editing, the only comment here is the preposition 'in' on line 276, maybe it should be 'to'… english is not my mother's tongue and my skills are deteriorating. .. well also maybe line 166 *markedly* would work better than *wildly*. Colorful language works better in reviews.
- On line 110 it can also be said that the entrapment rate used i models is often not as physical as they purport. The values rather express an imposed constraint on precisely the quantities the manuscript diagnoses.
- I had the impression that fewer color steps… i.e., a half dozen or less, and a logarithmic spacing of the histograms would make the shape of the distributions somewhat more clear.
- Finally (line 100) a manuscript that doesn't say something is out of scope, but rather says what is in scope by virtue of which other things are out of scope. Kudos