I found this article very interesting, and I have a couple of points for the authors to consider that I hope I have not expressed too incoherently.

First, I would like to push back against the language that features in the abstract and introduction stating that cloud and aerosol properties are averaged to 20 – 100 km resolution to reduce uncertainties. Averaging is not a strategy to deal with uncertainty. Averaging reduces random error, but any statistical estimation procedure can deal with such random errors, even if they are heteroskedastic across the samples/pixels. In fact, in the heteroskedastic case, simple averaging is not a good idea even if any subsequent inference is based entirely on areal averages. Beyond that, we know that remote sensing retrievals of cloud properties are not dominated by random error (e.g., radiometric noise). Instead, it is deterministic error due to algorithmic assumptions and finite resolution (i.e., errors are functions of the unknown cloud state) and spatially correlated errors in ancillary data that are the dominant sources of uncertainty. We would need to know how to model these deterministic errors with error covariances to propagate uncertainty. If we don't know how to model these errors, it doesn't matter whether we average or not, uncertainty is unspecified and the issue remains. So, the averaging is largely orthogonal to the matter of uncertainty and its propagation in inference. We have chosen to average; it is not a behavior that is prescribed by measurement limitations. Certainly, averaging doesn't meaningfully address any significant measurement limitations.

I suggest instead that conceptual simplification and practical issues of data size and computing power are the reason for averaging. As we have progressed in our understanding, we have moved from asking simple questions such as "What make more cloud?" to asking "What controls the intra-cell covariance of droplet concentration and liquid water path across closed-celled stratocumulus?" Answering the former question only requires coarse averages. As our theories grow in detail and subtlety we need to interpret our measurements with more nuance. This leads to a natural transition from only using coarse-resolution averages to analyzing the details of the spatial structure of cloud fields. I agree with the author's suggested direction; that there is much more to learn from snapshots when we interpret them correctly.

It appears to me that there is an assumed separation between the estimation of 'the rules of the game' and the knowledge that 'the rules are invariant' across a set of samples. To me, it is not clear that this is the case. When reading, I don't see a clear definition of which geophysical variables or properties we can use as evidence that 'the rules are invariant' and which we can use to determine the rules themselves (i.e., constrain processes).

To put this in more concrete terms, let's say that we want to perform a Bayesian inference of a vector of parameters of a microphysical parameterization, $\theta$ using a vector of observations $y$. We

need to consider the influence of the meteorological state, $\gamma$, on our observations so we must estimate the posterior distribution of these two sets of variables together:

$$p(\theta,\gamma|y) = \frac{p(y|\theta,\gamma)p(\theta,\gamma)}{p(y)} \tag{1}$$

What I believe is being assumed is that these kinds of estimation problems are separable so that

$$p(\theta,\gamma|y) \approx p(\theta|y_1)p(\gamma,y_2) = \frac{p(y_1|\theta)p(\theta)}{p(y_1)}\frac{p(y_2|\gamma)p(\gamma)}{p(y_2)} \tag{2}$$

where $y_1$ and $y_2$ are non-overlapping subsets of observations.

For example, when we perform a reanalysis, we estimate large-scale thermodynamic and dynamic properties such as winds, temperature and humidity using measurements of those same properties, $y_2$, but not measurements of cloud microphysics ($y_1$), and we do not jointly estimate the microphysical parameters, $\theta$. So, we obtain a maximum a posteriori estimate of the meteorological state (i.e., reanalysis):

$$\tilde{\gamma} = \underset{\gamma}{\operatorname{argmax}}\, p(\gamma|y_2) \tag{3}$$

Then, we use measurements of cloud properties, to estimate parameters of cloud processes conditioned on knowledge of the meteorological state from reanalysis:

$$p(\theta|y_1,\tilde{\gamma})\frac{p(y_1|\theta,\tilde{\gamma})p(\theta)}{p(y_1)} \tag{4}$$

In Eq. 4 we find the 'rules of the game', $\theta$, from observations under the assumption that the rules are invariant (fixed $\tilde{\gamma}$).

For example, for the stratocumulus case, it is stated that the inversion height is horizontally homogeneous, so the 'rules are invariant', and yet variation in the inversion height appears to be integral to the intra-cell variability as well from Fig. 2. Is there a clear way to identify this separability? Again, for the stratocumulus, do we know a priori that there are no drivers that operate at scales between the cellular scale and ~100 km? Or are we relying on observations (reanalysis?) that demonstrate a lack of variance at this range of scales? For the cold-air outbreak trajectory example, the timescales discussed only mentions the timescale of SST gradient. Could there not be meteorological changes that are significant at a timescale of ~12 hours associated with synoptic systems that cold-air outbreaks are often part of?

I think it would be great if the authors could be a bit more precise in how they would determine that the 'rules are invariant'. Eq. 4 seems to be assumed ubiquitously. For example, we assume we do not need to solve a data assimilation problem to calibrate a climate model. In other words, we assume that "we don't need to know the weather to project the climate", though I've yet to see any

evidence of this. Is this separability real or do we impose (assume?) a scale-break between the resolution of global reanalysis/climate model and the domain size of Large Eddy Simulations that is just an artifact of computational limitations? This is a critical assumption that also appears to underly the authors' arguments, so it would be great to get their opinion on it.

The authors' notion of how processes can be inferred from snapshots is more specific than my Eq. 4, arguing that, for Type 1 systems, there is information about a fast microphysical process ($\theta$) even when $y_1$ are effectively contemporaneous. Am I correct in understanding that ergodicity implies that we can interpret the droplet effective radius profile in cumulus or the cellular structure in closed-celled stratocumulus using a parcel model, rather than requiring a whole LES? If this is true, it becomes a lot simpler (especially computationally) to evaluate the likelihood, $p(y_1|\theta, \tilde{\gamma})$, and the corresponding posterior. If this is the case, it would be good for the authors to state it explicitly or if not, it would be good to have a more practical statement of how exactly ergodicity would simplify inference of processes. Even in the effective radius profiling example, there is a clear influence of thermodynamics through condensation rates etc., and again the issue of separability in the inference of a microphysical process arises. I agree that developing a deep understanding the mechanism for the apparent ergodicity is extremely important to justify this sort of observational interpretation.

The notion that processes can be extracted from Type 1 snapshots suggests to me that we might get more value from observing systems that provide high detail and accuracy in select conditions (at the expense of sparse sampling) rather than those that sample everything but with little detail or precision. Does this align with the authors' understanding? If so, it might be worth making a recommendation along those lines.

Perhaps I am misunderstanding, but I have some concerns about the Type 2 cases, where it is stated that they may be useful after careful stratification by meteorology. If drivers such as aerosol and 'meteorology' are correlated across snapshots, then stratification (or other statistical models and their counterfactuals) will produce biased estimates of how clouds respond to an aerosol driver under 'constant meteorology' (and vice versa). The assumption that including a variable as a covariate in a statistical model will control for its influence is a fallacy. For example, if the mechanism by which aerosol affects cloud fraction is changes in sub-cloud stability from precipitation suppression, then including a variable that correlates with any of the variables involved in this process as a covariate in a statistical model or as a stratification parameter will bias the apparent susceptibility (i.e., partial derivative) of cloud fraction to aerosol with respect to a true causal response. Statistical frameworks for analysing 'causality' can only untangle these

effects when processes are resolved by the measurements, i.e., not Type 2, and all relevant variables in the causal graph are observed.

The stratification approach seems to underpin the entirety of ACI analysis. It is unsurprising then that direct estimates of things like radiative forcing due to ACI from statistical counterfactuals seem highly disconnected from the actual large-scale behavior and processes, as recently found for liquid water path adjustments. To borrow the statistical terminology of medicine, cross-sectional statistics (like sets of snapshots) are not definitive but can motivate proper study design (e.g., randomization) to estimate effect sizes. Even in a longitudinal study (i.e., tracking clouds with geostationary satellites), there is difficulty in untangling causality.

I would suggest that direct estimation of 'processes' from statistical models built on Type 2 observations is not going to be robust, as 'drivers' will also be correlated with intermediate variables. Instead, I would advocate for an observation-constrained model-based counterfactual in which Eq. 4 (or better Eq. 1) is evaluated and then the calibrated model is used to compute a counterfactual. In other words, I am arguing that Type 2 cases do not provide a shortcut to access process understanding. I think it would be helpful for the authors to be a bit more precise about the conditions required for Type 2 cases to be helpful for process understanding in terms of controlling for the variation of slow processes across snapshots.

I strongly support the closing statement that existing measurement limitations should not get in the way of the refinement of conceptual thinking regarding how we can best extract understanding from measurements. In fact, I hope that refining our conceptual thinking will drive innovation in measurement. For example, currently satellite remote sensing measurements formulate their scientific accuracy requirements without any consideration of spatial error covariance (This is the cause of the uncertainty issue discussed first). If we show that process understanding comes from spatial patterns in snapshots, then the requirements should be set in terms of spatial error covariance. Confronting our existing algorithms with such a requirement will drive innovation.

The authors make the statement:

Line 401-402: "At high solar zenith angles (SZA) retrievals are more problematic but events that lie within the optimal SZA window (less than $65^{o}$; Grosvenor et al., 2018) will be valuable."

This statement that events within the SZA < 65 are optimal and therefore valuable for studying cloud processes is not consistent with the available evidence.

Here, the claim that SZA > 65 are insufficiently accurate is conflated with the claim that retrievals with SZA < 65 are sufficiently accurate. A reading of Grosvenor et al. 2018 and the references within reveals that the sufficiency of operational geostationary retrievals with SZA < 65 to be

‘valuable’ for studying the covariance of droplet concentration and liquid water path etc. has never been demonstrated. The conflation of these points is a widespread fallacy in the use of satellite remote sensing data to study aerosol cloud interactions, i.e., “we excluded the lowest-quality data so now what we have left are good quality” (not just less low-quality). See Loveridge & Di Girolamo (2024) for more discussion of this point.

I suggest that the authors simply follow the spirit of their closing statement and avoid distracting from their main point by discussing details of measurement performance. The main point of this paragraph, that measurements with wide field of view and high temporal frequency will be useful, has the same caveat as all measurements (sufficient accuracy) that are discussed in the article. I don’t think the authors should stress over justifying this particular type of measurement.

References:

Loveridge, J. R., & Di Girolamo, L. (2024). Do subsampling strategies reduce the confounding effect of errors in bispectral retrievals on estimates of aerosol cloud interactions? *Journal of Geophysical Research: Atmospheres*, 129, e2023JD040189. https://doi.org/10.1029/2023JD040189