REVIEWER 1

This paper assesses the usefulness of satellite-measured atmospheric temperature and humidity for convection prediction. The assessment is based on surrogate data as opposed to real measurements. Specifically, global reanalysis data are sampled according to the sampling pattern of a polar orbiting satellite, to mimic the retrievals of an infrared hyperspectral sensor, AIRS. One interesting aspect of this investigation is the use of a trajectory model, which was introduced in an earlier work (Kalmus 2019), to increase the spatiotemporal representativeness of the satellite measurements. The paper is logically organized and well written, providing sufficient technical information and clear descriptions of the results. I do have some concerns, as detailed below, on several aspects of the paper, including the design of the research, the method, and the interpretation of some results. I think this paper could add excellent contribution to the literature after these comments are addressed.

Thanks for carefully reviewing our paper and for your encouraging comments.

We respond point-by-point below but first summarise two points that merit particular attention.

Firstly, we provided more context and justification for our experimental design choices, in particular why we include all valid profiles even after convection has happened. Our planned uses involve predicting convection from thermodynamics. Predicting "does" versus "does not" convect cannot be reliably done by only learning from cases in which it does. In a new discussion paragraph, we describe our plans and use Bayesian framing to explain our choices. The argument is specific to our purposes though, we did follow one of your suggestions testing T and q statistics during precipitation. The results are shown in a new supplementary figure and are useful for testing some of our statements regarding processes.

Secondly you raise important issues related to sounder retrievals. We want to evaluate FCST without restricting results to a single sounder product. We've added text but tried to ensure that conclusions will not *only* apply to a particular AIRS product version. New discussion covers AIRS v7's non-analysis priors and potential trend biases. Then a new supplementary analysis studies the consequences of changes in vertical & horizontal resolution for our conclusions. The new analysis is general in its treatment of resolution, so that conclusions should be easily translated to other potential sounder products, but we briefly report results from actual AIRS data in the supplement too.

We hope that the additional text and supplementary analysis satisfies your concerns and once again, we appreciate your very thoughtful commentary.

Your review text is **bold black**, our commentary is magenta (this colour!) and any quoted text that is in the main paper is in quotation marks and is "dark green".

L30-35. Two points are provided as the motivation of this work: weather and climatology. These starting points probably need to be reflected on or revised. For the objective of improving weather forecast, since the trajectory relies on NWP-model modelled winds, how could this approach have any advantage over the data assimilation approach? For the objective of studying convection climatology, why not simply use the reanalysis data without reducing the sampling to match AIRS?

We have changed text to better justify. AIRS-FCST aims to allow study of climate trends with some independence from other methods. This is analogous to how e.g. trends in SSTs have been studied with buoys, ships, Argo floats & satellite products to increase our overall confidence in the real physical changes, even if each approach has its own uncertainties.

The new text includes:

"A common issue in reanalyses is that changes in the type of data assimilated result in discontinuities, which then cause trend biases. For ERA5, this has been demonstrated for snow cover (Urraca and Gobron, 2023) and upper-tropospheric temperatures (Shangguan et al., 2019). While ERA5 assimilates AIRS brightness temperatures, using AIRS retrievals alone avoids the reanalysis-specific concerns about changes in assimilated data for T and q, offering a complementary perspective on trends in the pre-convective atmosphere."

Using a single data source avoids one type of common issue that results in trend biases. All LEO sounders have sampling issues, so we need to include that factor to address our research goals:

"we take ERA5 T and q fields and convert them to AIRS time and space sampling and then use NWP trajectories to generate a time-resolved product for comparison with ERA5 outputs at later timesteps. We follow AIRS time-space sampling to ensure that our statistical results apply to future work using AIRS data."

And repeat the point for emphasis at the end:

"The primary motivation is to guide development of AIRS-FCST for climate studies, with a focus on the "FCST" component rather than issues related to any specific set of AIRS retrievals. Nevertheless, a fundamental limitation of any LEO-based product is the spatial sampling of the instrument, so AIRS spatial sampling effects are also studied."

With reference to NUCAPS, we mention some advantages,:

"Operationally, NUCAPS-FCST can provide users with information based on the latest satellite T and q fields sooner than if they waited for the next NWP forecast cycle. FCST can also provide complementary information since compared with NWP its results are less sensitive to convective parameterisations."

Based on present tests we expect to be able to get the NUCAPS-FCST fields into AWIPS II within ~1 hour of overpass.

L82. An important claim is made here about AIRS being advantageous for studying climate trends compared to reanalyses. This point needs to be better discussed, as one can easily come up with counterarguments. For example, given that the conventional retrievals typically take prior information including first guesses from analysis, it is not obvious to me that the retrieval products aren't subject to the same issues as reanalyses. A general comment is that I think the paper can provide better reasoning or more references to establish suitability of AIRS for studying climate trends. For instance, do you think the radiometric stability of AIRS together with its spectral information may facilitate detecting convection regime changes, taking advantage of their spectral signatures (e.g., Huang and Ramaswamy 2008, https://doi.org/10.1029/2008GL034859; Kahn et al. 2016, https://doi.org/10.1002/2016GL070263)? Or, may methods particularly designed for climate trending, such as the average-then-retrieve approach (e.g., Huang et al. 2010, https://doi.org/10.1029/2009JD012766; Kato et al. 2014, https://doi.org/10.1175/JCLI-D-13-00566.1) be of relevance here?

We agree with the reviewer that more finesse was needed in the supporting argument. We have added the following text to the introduction, which covers how the v7 retrieval first guesses use only the radiances as input:

“The AIRS v7 retrieval uses a neural network (NN) to generate its first guess, including T and q profiles. The NN inputs are AIRS measurements, but the NN training dataset was based on ECMWF forecast profiles (Milstein and Blackwell, 2016). The AIRS prior may therefore share common structural biases with reanalysis, but its trends should respond to radiances alone, since the NN is “expected to behave similarly throughout the whole mission, while model-based first guesses can show significant change in bias structure over mission duration due to model changes” (Yue et al., 2020).

Relevant issues with AIRS retrievals and the potential for climate trend studies will be further discussed in Section 4, but retrieval uncertainties are ignored by design since the purpose of this study is to isolate and quantify errors associated with our method of adding time resolution to AIRS T and q fields.”

Then in Section 4 we further discuss:

“Despite Yue et al. (2020)’s expectations that the AIRS neural network (NN) prior is unlikely to cause trend biases, machine learning algorithms can drift if the environment differs from their training set. Prior-related biases in any AIRS-derived product should therefore be independent of reanalysis trend biases, since ERA5 assimilates brightness temperature rather than AIRS retrievals. Therefore, AIRS-FCST will provide independent information to study climate trends. Furthermore, future AIRS product development could use an updated training set to remove environmental drift. Ongoing AIRS-FCST development work will investigate AIRS-related issues in more detail, while this paper has established a deeper understanding of FCST related issues.”

We considered discussion about AIRS product trends in general, but didn’t see an obvious place to enter those citations. With the thermodynamics we are targeting hourly dynamics, and the rest of the discussion touches on hazardous convective weather, which is not well-captured at AIRS overpass time. We think our work is substantially different from the citations you mention, author Kahn supports the choice to exclude these citations.

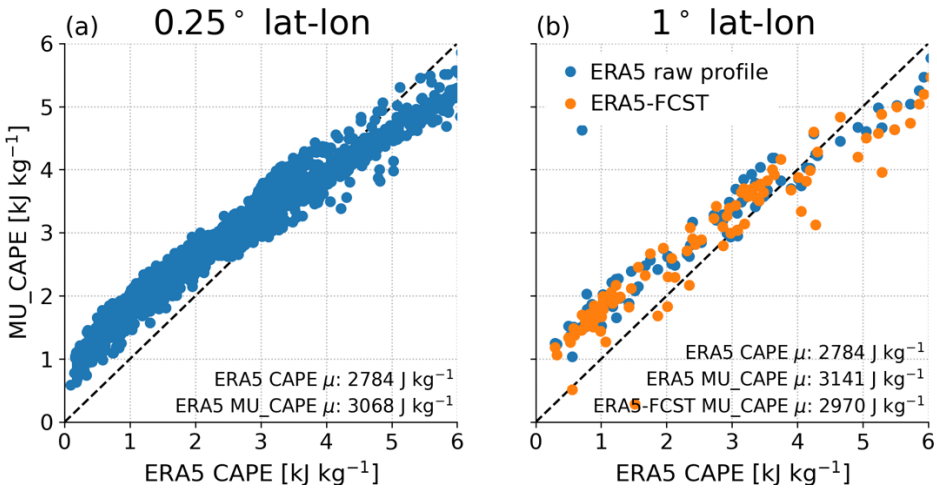
L129. A critical methodological question here is whether ERA5 profiles can appropriately represent the vertical resolution of AIRS retrieval or its ability in measuring such quantities as CAPE. I'm surprised that this important consideration is completely neglected. The paper would benefit from a proper discussion of this issue or an assessment of the impacts, for instance, by using the AIRS averaging kernels.

This is a very good point and we should provide readers with evidence. Rather than substantially lengthening the paper, we have added a supplementary analysis and refer to it in the main text via:

“The convective parameters are calculated from the final gridded fields at 30 hPa vertical resolution, and any profile comparisons are based on the same output. We expect our results to be robust to changes in vertical resolution between ERA5, AIRS L2Sup and the final outputs based on a series of resolution sensitivity tests for derived CAPE (Supplementary Figures 1—3, Supplementary Table 1).”

The supplementary analysis prefers CAPE-calculation sensitivity tests because we believe it sufficiently addresses the point of spatial resolution. An averaging kernel analysis would have been lengthy and difficult to explain to readers who’re unfamiliar with standard Bayesian retrieval frameworks.

We show how our results aren’t very sensitive to vertical or horizontal averaging, with figures such as a new Supplementary Figure 2, reproduced here:



ERA5 CAPE represents the result using the highest possible resolution, including 137 vertical levels. Associated supplementary text then goes into more detail to show how our *binning* of grid cells by CAPE means that our results are not sensitive to nonlinearities or mean differences in CAPE between different calculation methods. It is only if the relationship is not monotonic or if scatter is large that our results are sensitive to CAPE calculation method and vertical resolution.

The paper tries to distinguish between issues related to retrievals (AIRS *or* NUCAPS, although we mainly discuss AIRS) and the FCST procedure. We simply have to establish whether FCST is at all valid before we can justify any products using this technique, and we believe the present submission quantifies its performance in a suitable manner.

The consequences of retrieval errors will be investigated as we progress, but we report some preliminary statistics based on resampling AIRS v7 L2Sup onto the ERA5-FCST grid (1° horizontal, 30 hPa vertical) at the end of the new supplementary analysis.

L155/L319. What "neglected processes" are referred to here?

Where introduced (paragraph near original L155) we are now more explicit:

"The only way in which ERA5-FCST parcel T and q can be affected is via vertical motion and the associated adiabatic heating or cooling. We refer to diabatic processes such as radiation, surface fluxes and sub-grid convection as "neglected", even though they indirectly affect results since the NWP simulation that provides the winds for HYSPLIT includes these processes. Nevertheless, the neglected processes can greatly affect T and q profiles in ways that are not captured by FCST. Sub-grid convection in particular can rapidly transport heat and greatly change local profiles, but can still have a relatively small effect on the motion vectors once averaged over a large NWP grid cell. This is because the rising warm air within a grid cell is compensated by nearby descent."

L205. A relevant question of interest is how much the 1:30am/pm overpass times of AIRS limit the convection prediction. Or, what different times would be more useful? Can this study provide some insights?

We decided against adding content on this for three reasons:

1. The analysis we have doesn't add much beyond what's known, which is "it depends on your time and location", e.g. late day convection in the U.S. often starting near the Rockies then progressing eastward.

2. Our AIRS-FCST goal is to target climate trends in the past. Any suggestions on when to sample wouldn't help for the past, while for the later 2020s there are planned geostationary hyperspectral IR sensors for which it doesn't make sense to suggest an overpass time.
3. To say something new will take a *lot* of investigation and very likely more than the 9-month sample we have here.

We definitely intend to investigate the diurnal cycle in AIRS-FCST with its ~20 years of data though.

L292. The poor prediction of the temperature of the upper layers (fig. 7c) is surprising. Why?

We think that this is consistent with a point we make later – when there's a lot of intense convection going on, ERA5 transports a lot of heat into the upper troposphere. The most obvious feature to us in Fig. 7(c) and Fig 8(b,c) is that the point "cloud" is shifted to the left near the origin, i.e. high-level cooling in ERA5-FCST but not in ERA5. To prepare readers we have added an earlier comment:

"As the convection passes overhead after 0000 UTC, it is also notable that the upper troposphere from 100—400 hPa cools substantially more in Figure 6(c) compared with Figure 6(b). The weaker cooling in ERA5 may be explained by sub-grid convection pumping heat into upper levels as the storm passes."

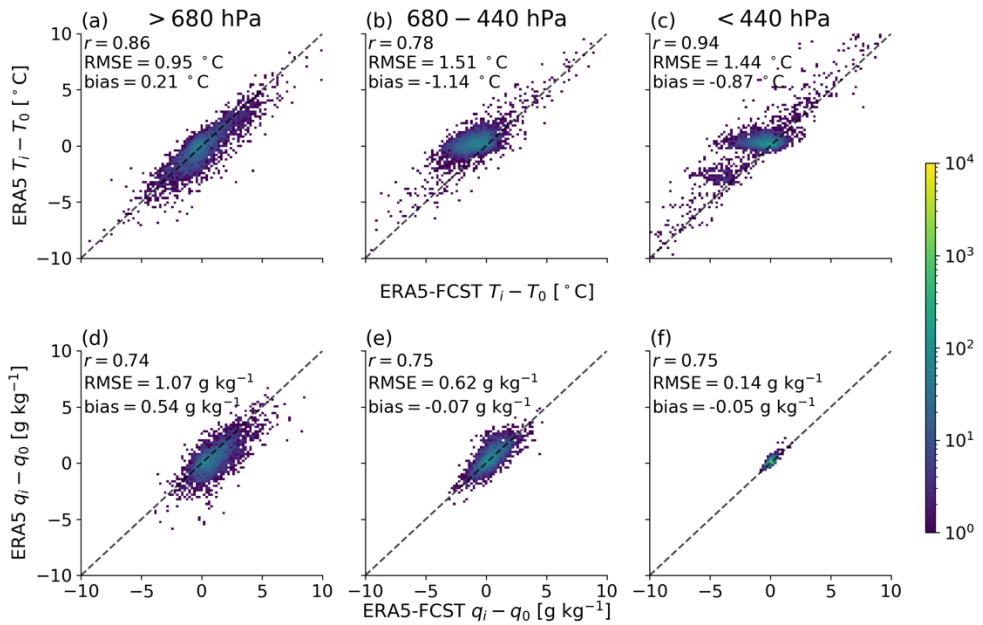
With this signpost we hope readers will now be able to interpret the comments later describing Figure 7. We have also added a call back to the Fig. 7 discussion:

"For middle and upper layers, the ERA5-FCST product projects larger temperature variation than occurs in ERA5, as was noted for Figure 6(b,c) after convection occurred."

In response to a comment by reviewer 2, we have added a new supplementary figure 7

"This mid- and upper-layer cooling in ERA5-FCST relative to ERA5 is indeed more common during JJA 2019 when convection peaks (Supplementary Figure 6), or during timesteps when precipitation exceeds the 99th percentile (Supplementary Figure 7)."

Supplementary Figure 7 is reproduced below, its caption points out how the mid- and upper-level bias described is larger in precipitating columns, which is consistent with our proposed mechanism.



L321. Is this really surprising since diabatic heating tends to be balanced by adiabatic motion at grid (or large) scales? And, again, clarify what's "neglected". Another philosophical question here is that there is equal amount (50%) of unexplained variance – this raises many questions:

how does this limit the usefulness of the prediction, and in what situations - for example, what regions or weather systems are missed?

We have added another sentence to explicitly call out what we think is happening:

"We expect that the unexplained variance is explained due to the contribution of the neglected diabatic processes, such as sub-grid convection, and to a lesser extent by differences between WRF versus ERA5 winds."

Combined with our other changes to your earlier comment including the explicit definition of "neglected", we think this covers the points.

L447. Following this reasoning, shouldn't the analysis and comparison be limited to times prior to convection?

That would be ideal, yes! In fact, one of our goals is to attempt to develop a "classifier" or similar that will tell us when convection is likely to occur. For our planned use cases, it doesn't make sense to study times prior to convection.

Nowcasting, e.g. NUCAPS-FCST – forecasters will have the thermodynamic fields and have the job of inferring things such as convective risk. *They do not know with certainty whether convection will happen*, so results based only on cases where convection happens would be difficult for them to use.

Climate, e.g. AIRS-FCST – for 2002—2013 we don't have a spatially complete, quality controlled & quantitative dataset of convective risk. For 2014—2020ish MRMS will give us that "climate quality" data, with consistent processing, sampling etc. Before MRMS we will therefore *only* have thermodynamics.

We have added text in Section 4 to try and describe, and use some Bayesian terminology to try and keep things precise and concise:

"The 2002—recent AIRS-FCST record of thermodynamics will be used with the MRMS surface radar (2014—recent) to relate the derived thermodynamics to convection. In a Bayesian sense, AIRS-FCST will provide P(thermodynamics) and to obtain our target of P(convection) we aim to derive P(convection|thermodynamics) using the combination of AIRS-FCST and MRMS. The proposed analysis is subtly different from previous work such as Kalmus et al. (2019), which studied thermodynamics in convective versus non-convective atmospheres and so reported results relevant to the inverse problem of P(thermodynamics|convection). We also emphasise that while the present study considered CAPE and CIN, this is a proof of concept that only considered a subset of potential thermodynamic properties."