nature portfolio

Peer Review File

A data-to-forecast machine learning system for global weather

Corresponding Author: Professor Hao Li

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This works presents a novel method for cycling a data assimilation system using only ML models, without a physical model or traditional DA methods. This is a unique system: it takes a background field from an ML forecast, combines the background with satellite observations via ML, and produces an analysis field for initializing forecasts with an ML model. The results are impressive and this work is an exciting step forward in the field of ML for weather prediction. The paper is well-written and the conclusions are generally supported by the results.

Major comments
1. What is the main reason for withholding conventional observations? What are the challenges?
2. Supp Fig 3b and SI lines 321-328: Why do the DA windows overlap? Does this mean that some observations are effectively used twice? In traditional DA, this is generally avoided as this can cause overfitting and ultimately degrade forecast skill. Can you comment on the reasoning behind this design choice, and if observations are assimilated more than once, can you explain whether this would cause any issues or overfitting?
3. Main text Section 3.4 needs a bit more explanation (either in the main text or SI) and I don't fully understand Figure 4. It's possible my misunderstanding is due to my lack of experience with satellite DA, but a bit more explanation here may be useful for other readers like me. It's possible my misunderstanding is due to my lack of experience with satellite DA, but a bit more explanation here may be useful for other readers like me. Specific comments related to this:
a. Can you clarify whether the perturbation was introduced directly to the brightness temperature? Was the observation perturbed for all channels from NOAA-20, or for several channels separately?
b. Related, I don't understand Figure 4: are these "increments" from three different perturbations (to different channels)?
c. I'm also not sure the phrase "analysis increment" is accurate here. For example, Figure 4 is not showing analysis increments in the traditional sense (analysis minus background) but rather the difference between two analysis fields (which are perturbed by a single observation). Is that the correct understanding?
d. Finally, SI lines 518-519 mention that this perturbation was applied across a 5x5 grid point. Does this mean that the "red dot" in Figure 4 should actually be a box with sides of length 1.25deg? Are there any other implications of this choice?
4. SI Section 6 needs clarity on difference between figs 13 and 14 and what the implication is. Specifically, SI lines 569-576 are unclear to me. It seems like this is suggesting that the first set of experiments shows how the analysis fields were degraded but the second set shows how subsequent forecasts were also degraded. But, I thought that backgrounds were not used in the second set of experiments? Can you expand on this a bit?

Minor comments
1. Line 38: integrates -> integrate
2. Lines 101-109: observation error correlation is also a major reason for only assimilating a small amount of available satellite data. Is this addressed in this method?
3. Lines 112-113: should maybe be "dynamical systems"?
4. Line 131 and elsewhere: "cyclic DA" should maybe be "cycled DA" or "cycling DA"
5. line 161: consider adding the satellite data preprocessing step to Figure 1?
6. Line 165 and SI section 2.1: can you comment on why the interpolation of observational data to the model grid is necessary? Is it too difficult to have the NN do this interpolation itself?

7. Line 167: what is meant by “data modalities”?
8. Line 202 “referred to as the control run” I couldn’t find any other references to a “control run” in the main text; maybe this phrase should be removed here. Instead, it could be defined only for SI section 6, where it is used. This could simply be replacing SI line 537 with “…was assessed based on experiments that assimilate the full set of satellite data, which will be referred to as the control run.”
9. Ext. Data Fig 2: Why do the 42h forecasts have lower error than the analyses?
10. Line 223: should be “higher at 850 hPa than 300 and 500…”
11. Line 235 “markedly enhanced the accuracy”: can you say anything about statistical significance here? If I understand correctly, the shading on Ext Data Fig 2 is the original data, not a statistical error bar.
12. Line 243: “the figure” can you clarify if this is ED Fig 2?
13. Line 248-250: I’m missing some intuition here. Does this imply that the background fields will be more accurate twice per day? If so, why doesn’t this translate to a more accurate analysis field as well?
14. Lines 264-265: this wording is a bit awkward. Maybe, “…it outperformed ECMWF HRES after a lead time of 2-8 days, depending on variable and pressure level.”
15. Line 274: analysis -> analyses
16. Line 326: “using substantially less…”
17. Figure 4 caption needs to be updated (it refers to the “second and fourth rows”.)
18. Lines 354-355: I don’t see the “larger increment values downstream of the wind fields” in Figure 4. The “increments” are mostly centered on the perturbation location, or lie slightly north of the observation (but the winds appear to be generally westerly.) Can you clarify this statement?
19. Ext Data Fig 3 caption: green -> black
20. SI line 72: spatialtemporal -> spatiotemporal
21. SI line 388: By “ground truth”, do you actually mean “reanalysis”?
22. SI line 391: “similar to the those” -> “similar to those”
23. SI Section 3.4: this section is not referred to in the main text. In SI line 451, the reference to “the main text Section 3.4” may be inaccurate. I would suggest removing this subsection or clarifying what additional information is provided in this section. Possibly the relevant sentences could be moved to SI Section 6 (data denial experiments).
24. SI line 500: “Typically” Do you mean “Typically for satellite radiance observations,” ?
25. SI line 510 should be “we followed”
26. SI Section 6: Can you clarify in the text that each satellite was denied separately? This is only mentioned in the figure captions.
27. Supp Figs 13-14: Are there any negative values? The caption of Fig 13 refers to “red shading” but there is no comparable e.g. blue shading for negative values. Also, can you add significance to these figures?
28. Supp Figure 13 caption: clarify that this is the set of experiments that used background forecasts. Conversely, Fig 14 caption should describe that these experiments did not use background forecasts.
29. SI Line 569: “extended Data Fig 14” should be “Fig 14”.
30. SI lines 644-645: “…reveals that incremental learning has a more substantial positive impact than does fine-tuning alone.” Based on Supp Figs 21-22, it seems like this is the opposite? Setting 2 has much lower errors than Setting 3, which would mean that fine-tuning is more impactful than incremental learning. Can you clarify if the figure labels are correct and I’m interpreting this accurately?
31. Throughout main text and SI: I would recommend replacing MSL with MSLP as this is more standard.

(Remarks on code availability)
A README for the code would be useful for the community.

Reviewer #2

(Remarks to the Author)
The manuscript “FuXi Weather: A data-to-forecast machine learning system for global weather” by Sun et al., 2025 is an example of a growing area of research in Machine Learning applied to Weather Prediction (MLWP). Current MLWP models need to use initial conditions from traditional NWP systems (operational or reanalysis) to start their forecasts, which makes them effectively equivalent to advanced, 4-dimentional (space plus time) post-processing tools. It is thus of interest to see whether it is possible to extend the MLWP paradigm to include an autonomous data assimilation cycle and thus closing the circle from observations to forecasts.
The system proposed in this manuscript would still not be completely independent of external (i.e., ECMWF ERA5) reanalysis products as it uses the FuXi model to cycle the assimilation and run forecasts and FuXi is a model trained on ERA5 fields. Additionally the Authors explicitly state: “End-to-end training of FuXi Weather optimizes both the analysis and forecasts using ECMWF ERA5 reanalysis data at 0.25. resolution as the reference.” But it would certainly be a step in that direction.
Unfortunately, the paper does not convince me that the Authors have managed to achieve even this more limited objective. Specifically, the Authors evaluate the accuracy of the analysis fields produced by FuXi Weather by comparing their errors with those of 42-hours FuXi forecasts started from ERA5 reanalysis fields and verified against ERA5 reanalyses. Looking at Extended Data Fig. 2, these errors look comparable. My interpretation is that the analyses produced by FuXi Weather are about as accurate as 2 days forecasts produced starting the model integration from ERA5 reanalysis. This is a starting point for further research, but hardly a result worth discussing on its own merits.
Further, when the Authors compare forecast results from FuXi Weather with those from ECMWF IFS, they are mixing the effects of two separate components, the assimilation system and the forecast model. If they want to evaluate how good the assimilation system of FuXi Weather is, then they need to use the same forecast model (in this case FuXi) started from a separate set of initial conditions (eg, ERA5) as a control. Otherwise the different characteristics of the forecast models used

in the comparison muddy the conclusions. This is especially the case with MLWP models like FuXi which are known to suffer from increasing smoothness with forecast lead time and thus can game standard performance metrics.
More detailed comments are provided in the attached annotated version of the manuscript.
From my perspective, this is a potentially interesting but very preliminary investigation and certainly at too early a stage to be considered for publication.

(Remarks on code availability)

Reviewer #3

[Editorial Note: the attachments are displayed at the end of the file]

(Remarks to the Author)
The authors present an evolution of their machine-learning (ML) based forecasting system emulator Fuxi that is assimilating selected satellite based observations to create initial conditions for forecasts instead of using external analyses for initialisation. The data assimilation replacement method, called Fuxi-DA, is combined with an enhanced Fuxi forecast emulator to create the end-to-end Fuxi-Weather system that can be cycled like traditional, numerical methods based analysis-forecast systems. The output is assessed using standard forecast error metrics, and a specific focus is applied to a region in central Africa with sparse observational data coverage to test the robustness of the method.
The paper is complete regarding the explanation of methods, data, experiments and evaluation. The Fuxi developers build on an extensive record of ML applied to weather forecasting, and they exhibit deep knowledge of the methodological frameworks applied in the numerical weather prediction (NWP) community. They have also published their results before and made code available for others to try - which is highly commendable. The paper is very well written.
In the attached file, I have raised several points that will need addressing before publication. My greatest concern is the lack of evidence of statistical significance of forecast performance when compared against other data. More information is needed on the use of observational data in the data-assimilation process. I also wonder about the future need of data assimilation as an intermediate step between observational data records and forecast generation. Lastly, while I appreciate a special focus on central Africa for the reasons explained in the paper, the evaluation of performance for this case is not overly convincing.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors did a lot of work clarifying parts of the manuscript based on reviewer suggestions, which I appreciate. I have a few remaining comments.

Comments
1. Introduction and conclusion: It could be worth emphasizing that this is the first ML system that can cycle in a DA-like system for weather forecasting (to my knowledge, at least.) Specifically, this is the only system that I know of that can produce forecast fields that feed back into the system without blowing up. This, I think, is a major achievement. For example, Slivinski et al (2025) found that most deterministic ML models blow up within a few weeks of cycling in a traditional DA system.
2. Lines 160-161 “Furthermore, FuXi Weather’s development costs are considerably lower than traditional NWP systems.” This is not strictly true. FuXi Weather’s development relies heavily on ERA5, which built upon decades of NWP and reanalysis development. This is an important technicality, since many ML papers tend to gloss over this fact. With the exception of the AI-DOP techniques that learn directly from observations without reanalyses, every ML weather model is trained on a dataset that took years to develop and produce, and was built upon decades of knowledge.
3. Line 208: “a 1-year data” should be maybe “one year of data”
4. Lines 213-214: Is the variant of FuXi-DA without background forecasts discussed in the supplementary information? If not, I would consider adding it. Is this essentially direct-from-observation prediction?
5. Extended Data Fig 2: maybe replace “FuXi analyses” with “FuXi Weather analyses” to make the distinction between the two systems more clear?

6. Extended data Figs 5-6: I see that the authors added precipitation per a reviewer comment. However, there are some major issues with precipitation in ERA5 especially in the Tropics (Lavers et al 2022), such that this comparison is not very meaningful. I would urge the authors to consider the reviewer's other suggestion, to compare to observations; even a gridded observational product like GPCP (https://climatedataguide.ucar.edu/climate-data/gpcp-monthly-global-precipitation-climatology-project) or IMERG (https://gpm.nasa.gov/data/imerg) would help provide context here.
7. I still don't understand the reasoning behind using overlapping DA windows. Why was this design choice made? Can you articulate the motivation behind this design choice in the paper? Also, I would recommend explicitly stating in the supplementary information that some observations are used more than once, if that is the case.
8. Supplementary Info line 538: "a small perturbation at a single grid point." If I understand correctly, this is inaccurate, since the perturbation is in fact over several grid points, correct?
9. Captions of supp figs 18-22 and related description: May want to clarify that these are differences of RMSEs, as opposed to RMS differences, and define the direction of the difference. It would be helpful to add the description of what red vs blue means in the figure captions as well.

References:

Slivinski, Laura C., et al. "Assimilating observed surface pressure into ML weather prediction models." Geophysical Research Letters 52.6 (2025): e2024GL114396.

Huffman, George J., et al. "Improving the global precipitation record: GPCP version 2.1." Geophysical Research Letters 36.17 (2009).

Kirschbaum, D.B.; Huffman, G.; Adler, R.F.; Braun, S.; Garrett, K.; Jones, E.; McNally, A.; Skofronick-Jackson, G.; Stocker, E.; Wu, H.; et al. NASA's Remotely Sensed Precipitation: A Reservoir for Applications Users. Bull. Am. Meteorol. Soc. 2017, 98, 1169–1184.

Lavers, D.A., Simmons, A., Vamborg, F. & Rodwell, M.J. (2022) An evaluation of ERA5 precipitation for climate monitoring. Quarterly Journal of the Royal Meteorological Society, 148(748) 3124–3137. Available from: https://doi.org/10.1002/qj.4351

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
I wish to thank the Authors for their detailed responses and the time and effort devoted to run the additional experiments and produce the new diagnostics.
Unfortunately, I am still not convinced that the manuscript is ready for publication.
My main concern, among others, was that Fuxi-DA produces analyses that when used to initialise a forecast model generate forecasts that have ~2 days lower skill than those initialised with ERA5 initial conditions. To put things into perspective, a 2-day forecast skill drop is what we would see in a state of the art NWP DA system if **all** satellite observations were withheld. Bear in mind that satellite observations constitute ~95% of all assimilated observations in current operational NWP.
The Authors have two answers to this: 1) This is comparable to the state of the art in ML-driven DA, citing Allen et al. 2025 (https://www.nature.com/articles/s41586-025-08897-0); 2) The amount of observations assimilated by Fuxi-DA is smaller than that used in ERA5, especially considering the absence of conventional observations.
My perspective is that even though results of Allen et al., 2025, were not particularly convincing, results presented here do not show a significant improvement on those.
My suggestion to the Authors is to further develop their system by suitably extending the range of observations used, and only at that point it would be clearer whether end-to-end ML DA is a serious candidate for operational/reanalysis DA.

(Remarks on code availability)

Reviewer #3

[Editorial Note: the attachments are displayed at the end of the file]

(Remarks to the Author)
Please see attached file.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I appreciate the authors responses and modifications to the paper, and I think the paper is nearly ready for publication. I have only a few minor comments remaining.

1. Lines 161-163: consider including the Slivinski et al (2025) citation here.
2. Lines 395-396: consider replacing “indicating higher accuracy” with “relative to IMERG”.
3. Supplement: I think the added sentences on lines 94-101 regarding the 8h window would be more natural in the DA section instead (e.g. line 348).
4. Supplement lines 342-345: This is inconsistent with the statement that the assimilation window is 8h. These lines describe a window that is 7h long.
5. Supplement, added Section 9: I appreciate the authors added this section. However, my previous comment was more a request for the technical details of this “direct from obs” experiment. Was a different model trained for the experiments without background fields? Does the architecture of that model differ from the one with background fields? Or was the same model used, but with fewer inputs than it was trained to take? Maybe I missed it, but it seems like these details would be useful for interpretation of the results that background fields are helpful for forecast accuracy.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

(Remarks on code availability)

Review of “FuXi Weather: A data-to-forecast machine learning system for global weather”

Reply to Reviewer #1

We appreciate the reviewer’s meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #1 (Remarks to the Author):

This work presents a novel method for cycling a data assimilation system using only ML models, without a physical model or traditional DA methods. This is a unique system: it takes a background field from an ML forecast, combines the background with satellite observations via ML, and produces an analysis field for initializing forecasts with an ML model. The results are impressive and this work is an exciting step forward in the field of ML for weather prediction. The paper is well-written and the conclusions are generally supported by the results.

Major comments

1. *What is the main reason for withholding conventional observations? What are the challenges?*

Response:

The primary objective of this study is to develop a cycling data assimilation and forecasting system that relies exclusively on machine learning models. Specifically, we investigate the feasibility of achieving this goal using only satellite data. We agree with the reviewer that incorporating conventional observations, such as those from weather stations and radiosondes, poses additional challenges due to issues of spatial and temporal sparsity, inhomogeneity, and varying data quality. These challenges necessitate additional processing steps, including rigorous quality control. As part of our future efforts to enhance FuXi Weather, we plan to integrate additional satellite datasets and conventional observations. To improve clarity, we have revised the "**Discussion**" section accordingly (lines 457–463).

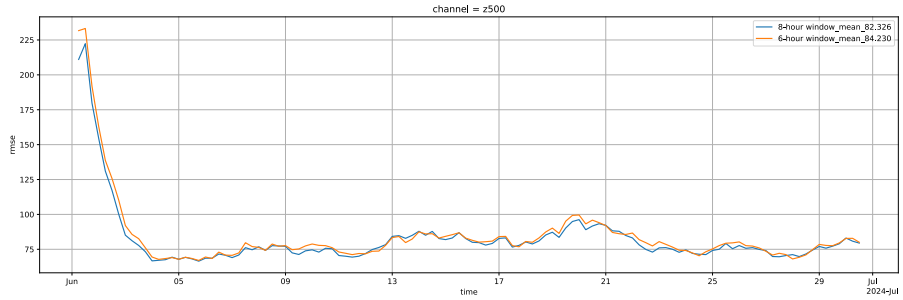
2. *Supp Fig 3b and SI lines 321-328: Why do the DA windows overlap? Does this mean that some observations are effectively used twice? In traditional DA, this is generally avoided as this can cause overfitting and ultimately degrade forecast skill. Can you comment on the reasoning behind this design choice, and if observations are assimilated more than once, can you explain whether this would cause any issues or overfitting?*

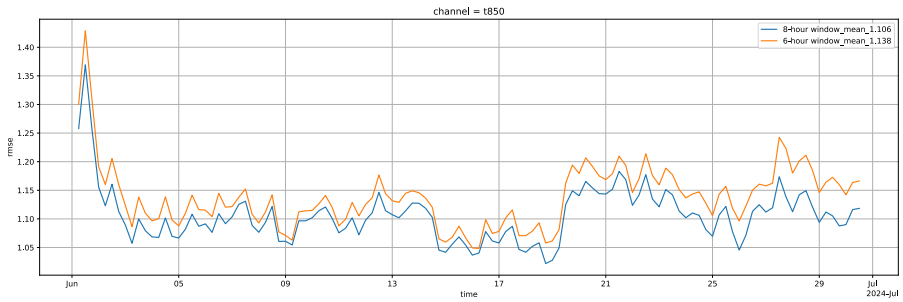
Response:

We thank the reviewer for this thoughtful comment. While overlapping DA windows are generally avoided in traditional DA frameworks, our experimental results suggest that minor overlaps in a ML-based DA framework do not degrade performance. For instance, when retraining FuXi-DA using a non-overlapping 6-hour DA window and evaluating its performance over a one-month period, we observed RMSE values comparable to those obtained with an 8-hour DA window configuration. As shown in the figures below, the differences in RMSE are marginal: for Z500, the RMSE values are 82.326 $m^2 s^{-2}$ (8-hour window) versus 84.230 $m^2 s^{-2}$ (6-hour window); for T850, the RMSE values are 1.106 K and 1.138 K, respectively. Given the negligible impact on performance and the considerable computational cost associated with retraining, we did not pursue further investigations.

We hypothesize that the reduced sensitivity of ML-based DA systems to DA window overlap, in contrast to traditional approaches, may arise from their ability to filter out processes associated with rapid early-time error growth in physical models, as discussed by Selz & Craig (2023, GRL). However, a comprehensive exploration of this mechanism is beyond the scope of the present study.

We have also included a corresponding statement (lines 341-344) in Section 2.2 of the Supplementary Information.





Reference:

Selz, T., & Craig, G. C. (2023). Can artificial intelligence-based weather prediction models simulate the butterfly effect? Geophysical Research Letters, 50, e2023GL105747. https://doi.org/10.1029/2023GL105747

3. *Main text Section 3.4 needs a bit more explanation (either in the main text or SI) and I don't fully understand Figure 4. It's possible my misunderstanding is due to my lack of experience with satellite DA, but a bit more explanation here may be useful for other readers like me. It's possible my misunderstanding is due to my lack of experience with satellite DA, but a bit more explanation here may be useful for other readers like me. Specific comments related to this:*

   a. Can you clarify whether the perturbation was introduced directly to the brightness temperature? Was the observation perturbed for all channels from NOAA-20, or for several channels separately?

   b. Related, I don't understand Figure 4: are these "increments" from three different perturbations (to different channels)?

   c. I'm also not sure the phrase "analysis increment" is accurate here. For example, Figure 4 is not showing analysis increments in the traditional sense (analysis minus background) but rather the difference between two analysis fields (which are perturbed by a single observation). Is that the correct understanding?

   d. Finally, SI lines 518-519 mention that this perturbation was applied across a 5x5 grid point. Does this mean that the "red dot" in Figure 4 should actually be a box with sides of length 1.25deg? Are there any other implications of this choice?

Response:

a. Yes, the perturbation was introduced directly to the brightness temperature, with only one channel perturbed in each experiment. To improve clarity, we have revised the corresponding sentences in Section 3.4 of the main text (lines 399–400 and 408–409).

b. Yes, these increments result from three separate perturbations applied to different channels. To enhance clarity, we have revised the sentence in lines 410-412.

c. We thank the reviewer for pointing out that the term "*analysis increment*" is inaccurate in this context. Accordingly, we have replaced it with a more appropriate expression. Specifically, we have changed the title of Section 3.4 to "*Physical consistency of* ***analysis changes***". Related revisions have been made to several sentences within Section 3.4 of the main text (lines 398, 402, 411–413, and 426), as well as in the caption of Figure 4. Corresponding revisions have also been made in Section 5.1 of the Supplementary Information (lines 511 and 539).

d. Yes, the perturbation was applied over a 5x5 grid point area. In traditional DA, data thinning is usually performed, but we did not apply thinning in our ML-based DA. As a result, observations remain relatively dense, leading to strong spatial correlations among them. Therefore, perturbing a single point may not effectively capture realistic observation error structures. By perturbing a 5×5 grid area, we ensure that the resulting impacts are more physically meaningful and better reflect realistic spatial error correlations. To clarify this methodological detail, we have added an explanatory statement in Section 5.1 of the Supplementary Information (lines 545–549).

4. *SI Section 6 needs clarity on difference between figs 13 and 14 and what the implication is. Specifically, SI lines 569-576 are unclear to me. It seems like this is suggesting that the first set of experiments shows how the analysis fields were degraded but the second set shows how subsequent forecasts were also degraded. But, I thought that backgrounds were not used in the second set of experiments? Can you expand on this a bit?*

Response:

Thank you for your helpful comment regarding the clarity of our explanation. We have revised the manuscript to better distinguish the two sets of experiments based on their use of background fields.

In the first set of experiments, all runs including those that exclude specific satellite data during DA, use the same background fields—specifically, the 6-hour forecasts from the control run. In contrast, the second set of experiments involves cycling DA and forecasts while excluding one satellite dataset in each case. As a result, the background fields vary between experiments, as they are generated independently from each respective cycling run rather than being fixed.

To clarify this difference, we have added a sentence to Section 6 of the Supplementary Information (lines 580-583). Additionally, Figures 13 and 14 have also been renumbered as Figures 14 and 15 in the update Supplementary Information.

Minor comments

1. *Line 38: integrates -> integrate*

Response: Thanks for highlighting this typo, we have fixed it.

2. *Lines 101-109: observation error correlation is also a major reason for only assimilating a small amount of available satellite data. Is this addressed in this method?*

Response:

We thank the reviewer for the insightful comment regarding observation error correlation as a major constraint on the assimilation of available satellite data. Our study addresses this issue in two ways. First, the observational dataset used in our experiments is relatively limited, consisting of only five microwave instruments from three polar-orbiting satellites, along with radio occultation data. Second, while our machine learning approach does not explicitly account for observation error correlations, this can be considered an inherent advantage of the method. The data-driven training process implicitly accounts for such correlations without requiring the specification of error covariance matrices.

To clarify this point, we have added a sentence to the "**Introduction**" of the main text (lines 105-106) and revised the relevant sentences in the "**Discussion**" of the main text (lines 455-462).

3. *Lines 112-113: should maybe be "dynamical systems"?*

Response: Thanks for highlighting this typo, we have replaced "*dynamic systems*" with "*dynamical systems*".

4. *Line 131 and elsewhere: "cyclic DA" should maybe be "cycled DA" or "cycling DA"*

Response: As suggested by the reviewer, we have replaced "*cyclic DA*" with "*cycling DA*" in the paper.

5. *line 161: consider adding the satellite data preprocessing step to Figure 1?*

Response:

We appreciate the reviewer's constructive suggestion. While we recognize the value of including the satellite data preprocessing step into Figure 1, this information is already thoroughly described in Section 2.1 of the Supplementary Information. To enhance clarity in the main text, we have revised the corresponding sentence in Section 2 (line 165).

6. *Line 165 and SI section 2.1: can you comment on why the interpolation of observational data to the model grid is necessary? Is it too difficult to have the NN do this interpolation itself?*

Response:

Thank you for your suggestion. To clarify the rationale for interpolating observational data to the model grid, we have revised the relevant sentence in the main text (lines 169–173).

7. Line 167: what is meant by "data modalities"?

Response:

In machine learning, the term "data modality" refers to a distinct form or method by which data are generated or represented. Meteorological observations from various sources constitute different modalities, as they capture diverse physical characteristics (e.g., visual, infrared, or point-based measurements) and originate from different sensing platforms (e.g., satellite-based versus ground-based instruments). To enhance clarity and avoid potential ambiguity, we have replaced the term "*data modalities*" with "*observation types*" in line 173 of the main text.

8. Line 202 "referred to as the control run" I couldn't find any other references to a "control run" in the main text; maybe this phrase should be removed here. Instead, it could be defined only for SI section 6, where it is used. This could simply be replacing SI line 537 with "…was assessed based on experiments that assimilate the full set of satellite data, which will be referred to as the control run."

Response:

Following the reviewer's suggestion, we have removed the phrase "*referred to as the control run*" from Section 3 of the main text.

Additionally, we have revised the original sentence in Section 6 of the Supplementary Information (lines 567-569) for improved clarity.

9. Ext. Data Fig 2: Why do the 42h forecasts have lower error than the analyses?

Response:

In Extended Data Figure 2, the analyses represent FuXi Weather analyses, while the 42-hour forecasts correspond to FuXi forecasts initialized using ERA5. Since FuXi-DA was trained using ERA5 as the reference dataset, the FuXi Weather analyses are not expected to surpass the accuracy of ERA5. Furthermore, FuXi-DA assimilates only a limited subset of satellite observations. As a result, the FuXi Weather analyses (incorporating background forecasts) show errors comparable to those of the 42-hour FuXi forecasts, with differences varying by variable and pressure level.

To improve clarity, we have revised the caption of Extended Data Fig. 2 accordingly.

10. Line 223: should be "higher at 850 hPa than 300 and 500…"

Response: Thanks for highlighting this typo, we have changed the word from "*with*" to "*than*".

11. Line 235 "markedly enhanced the accuracy": can you say anything about statistical significance here? If I understand correctly, the shading on Ext Data Fig 2 is the original data, not a statistical error bar.

Response:

We appreciate the reviewer's comment about statistical significance. In Extended Data Figure 2, the shaded areas represent the original data, while the solid lines indicate 12-point moving averages, as specified in the figure caption. Following the reviewer's suggestion, we conducted statistical significance tests comparing FuXi Weather analyses with and without background forecasts. The results show that incorporating background forecasts leads to statistically significant improvements in analysis accuracy.

To reflect these findings more clearly, we have revised the relevant sentence in the main text (lines 247-249).

12. Line 243: "the figure" can you clarify if this is ED Fig 2?

Response: We confirm that "*the figure*" indeed refers to Extended Data Fig.2. To eliminate any ambiguity, we have revised the text to explicitly state "*Extended Data Figure 2*" in place of the original phrasing.

13. Line 248-250: I'm missing some intuition here. Does this imply that the background fields will be more accurate twice per day? If so, why doesn't this translate to a more accurate analysis field as well?

Response:

We thank the reviewer for raising this important question regarding temporal variations in analysis accuracy. The relatively consistent performance of the FuXi Weather analyses across different initialization times can be attributed to two key aspects of our system design. First, the use of a fixed 8-hour assimilation window ensures temporal stability in the DA process. Second, the background forecasts used for DA are generated from previous FuXi Weather analyses rather than ERA5, resulting in more stable

analysis fields that are less sensitive to temporal variability than those initialized directly from ERA5.

To improve clarity, we have revised the corresponding sentence in the main text (lines 260-263).

14. Lines 264-265: this wording is a bit awkward. Maybe, "…it outperformed ECMWF HRES after a lead time of 2-8 days, depending on variable and pressure level."

Response: Following the reviewer's suggestion, we have revised the wording in the main text (lines 288-289) to improve clarity.

15. Line 274: analysis -> analyses

Response: Thanks for highlighting this typo, we have changed the word from "*analysis*" to "*analyses*".

16. Line 326: "using substantially less…"

Response: Thanks for highlighting this issue, we have removed the word "*an*".

17. Figure 4 caption needs to be updated (it refers to the "second and fourth rows".)

Response: We thank the reviewer for pointing out this inaccuracy in the figure caption. The redundant word "*fourth*" has been removed from the caption of Figure 4.

18. Lines 354-355: I don't see the "larger increment values downstream of the wind fields" in Figure 4. The "increments" are mostly centered on the perturbation location, or lie slightly north of the observation (but the winds appear to be generally westerly.) Can you clarify this statement?

Response: We sincerely appreciate the reviewer's careful analysis of this point. Upon re-examination of Figure 4, we agree that our initial characterization of the increment patterns requires clarification. The increments are mainly localized near the perturbation location, with a secondary region of moderate increments extending eastward in alignment with the prevailing westerly winds. This pattern contrasts with our earlier description, which emphasized larger downstream values.

To more accurately reflect these observations, we have revised the text in lines 421-423 of the main text.

19. Ext Data Fig 3 caption: green -> black

Response: We appreciate the reviewer for identifying this error in the caption of Extended Data Fig 3. We have corrected the color reference in the figure caption by replacing "*green*" with "*black*".

20. SI line 72: spatialtemporal -> spatiotemporal

Response: We have corrected the word from "*spatialtemporal*" to "*spatiotemporal*".

21. SI line 388: By "ground truth", do you actually mean "reanalysis"?

Response: We confirm that the term "*ground truth*" refers to the ERA5 reanalysis data. To avoid potential confusion, We have replaced "*ground truth*" with "*ERA5 reanalysis data*" in the Supplementary Information (line 403).

22. SI line 391: "similar to the those" -> "similar to those"

Response: Thanks for highlighting this typo. We have removed the redundant word "*the*" in the Supplementary Information (line 407).

23. SI Section 3.4: this section is not referred to in the main text. In SI line 451, the reference to "the main text Section 3.4" may be inaccurate. I would suggest removing this subsection or clarifying what additional information is provided in this section. Possibly the relevant sentences could be moved to SI Section 6 (data denial experiments).

Response:

We are grateful for the reviewer's observation regarding the inaccurate reference to "*the main text Section 3.4*", which should have referred to "*SI Section 6*". Following the reviewer's suggestion, we have removed this subsection from the Supplementary Information.

24. SI line 500: "Typically" Do you mean "Typically for satellite radiance observations,"?

Response: We thank the reviewer for this helpful clarification. As suggested, we have revised the relevant text in lines 526-527 in the Supplementary Information.

25. SI line 510 should be "we followed"

Response: As per the reviewer's suggestion, we have revised the phrasing to "*we followed*" in line 536 of the Supplementary Information.

26. SI Section 6: Can you clarify in the text that each satellite was denied separately? This is only mentioned in the figure captions.

Response: Following the reviewer's suggestion, we have added an additional sentence to improve clarity in line 585 of the Supplementary Information.

27. Supp Figs 13-14: Are there any negative values? The caption of Fig 13 refers to "red shading" but there is no comparable e.g. blue shading for negative values. Also, can you add significance to these figures?

Response: As suggested by the reviewer, we have updated the figures to include blue shading to represent negative values. We have also performed significance testing and used X symbols to denote where the resulting RMSE values are not statistically significant. The captions for Supplementary Figures 13 and 14 have been revised accordingly. These figures have also been renumbered as Figures 14 and 15 in the updated Supplementary Information.

28. Supp Figure 13 caption: clarify that this is the set of experiments that used background forecasts. Conversely, Fig 14 caption should describe that these experiments did not use background forecasts.

Response:

We appreciate the reviewer's valuable suggestion to clarify the methodological distinctions between Supplementary Figures 13 and 14.

In the caption of Supplementary Figure 13 (renumbered as Supplementary Figure 14), we have added the following text: "*Each experiment uses 6-hour control forecast backgrounds at analysis times, excluding one satellite data type during DA.*"

In the caption of Supplementary Figure 14 (renumbered as Supplementary Figure 15), we have included the text: "*Each experiment uses cycled background fields that evolve differently depending on the excluded satellite data.*"

29. SI Line 569: "extended Data Fig 14" should be "Fig 14".

Response: We thank the reviewer for highlighting this issue. And we have fixed it.

30. SI lines 644-645: "…reveals that incremental learning has a more substantial positive impact than does fine-tuning alone." Based on Supp Figs 21-22, it seems like this is the opposite? Setting 2 has much lower errors than Setting 3, which would mean that fine-tuning is more impactful than incremental learning. Can you clarify if the figure labels are correct and I'm interpreting this accurately?

Response:

We sincerely thank the reviewer for this careful examination. We confirm that the original statement was indeed the opposite. The relevant sentence (lines 712-713) in the Supplementary Information has been revised accordingly.

31. Throughout main text and SI: I would recommend replacing MSL with MSLP as this is more standard.

Response: As suggested by the reviewer, we have replaced "*MSL*" with "*MSLP*" throughout the main text and SI.

Reviewer #1 (Remarks on code availability):

A README for the code would be useful for the community.

Response: Following the reviewer's suggestion, we have uploaded the README file for the code to the Google Drive folder (https://drive.google.com/drive/folders/1q9xxKGqmR9tpjVd9R11OJ-eKf1nU10T-?usp=sharing_eil_se_dm&ts=6800f61c).

Reply to Reviewer #2

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #2 (Remarks to the Author):

The manuscript "FuXi Weather: A data-to-forecast machine learning system for global weather" by Sun et al., 2025 is an example of a growing area of research in Machine Learning applied to Weather Prediction (MLWP). Current MLWP models need to use initial conditions from traditional NWP systems (operational or reanalysis) to start their forecasts, which makes them effectively equivalent to advanced, 4-dimentional (space plus time) post-processing tools. It is thus of interest to see whether it is possible to extend the MLWP paradigm to include an autonomous data assimilation cycle and thus closing the circle from observations to forecasts.

Response:

We sincerely appreciate the reviewer's thoughtful and constructive comments, which have significantly contributed to improving the quality of our manuscript. We address the key concerns below.

The system proposed in this manuscript would still not be completely independent of external (i.e., ECMWF ERA5) reanalysis products as it uses the FuXi model to cycle the assimilation and run forecasts and FuXi is a model trained on ERA5 fields. Additionally the Authors explicitly state: "End-to-end training of FuXi Weather optimizes both the analysis and forecasts using ECMWF ERA5 reanalysis data at 0.25. resolution as the reference." But it would certainly be a step in that direction. Unfortunately, the paper does not convince me that the Authors have managed to achieve even this more limited objective.

Response:

We agree with the reviewer that the FuXi Weather system is not completely independent of ERA5 during the training phase. However, during the *inference phase*, FuXi Weather operates without relying on external NWP products. To enhance clarity, we have revised the relevant sentences in lines 192-196 of the main text to include this caveat.

Specifically, the Authors evaluate the accuracy of the analysis fields produced by FuXi Weather by comparing their errors with those of 42-hours FuXi forecasts started from ERA5 reanalysis fields and verified against ERA5 reanalyses. Looking at Extended Data Fig. 2, these errors look comparable. My interpretation is that the analyses produced by FuXi Weather are about as accurate as 2 days forecasts produced starting

the model integration from ERA5 reanalysis. This is a starting point for further research, but hardly a result worth discussing on its own merits.

Response:

Regarding the accuracy of analysis fields generated by FuXi Weather, achieving performance comparable to 42-hour forecasts represents a significant advancement in data-driven weather prediction. FuXi Weather is initialized with zero fields and operates in a cycling mode, performing DA and forecasting every 6 hours over one year. To our knowledge, the only other system currently pursuing a similar approach is that of Allen et al. (2025) (https://www.nature.com/articles/s41586-025-08897-0), which generates analysis fields with accuracy comparable to 2-day forecasts from ECMWF HRES.

- To provide a more comprehensive assessment, we have added Section 7 to the Supplementary Information, which compares the analysis activity (defined as the standard deviation of analysis anomalies relative to climatological means) and mean bias error (as described in SI Section 4) between FuXi Weather and ECMWF HRES. The results demonstrate strong consistency with ECMWF HRES in terms of analysis activity.
- It is worth noting that, in contrast to the extensive range of observational data used in generating ERA5, FuXi Weather currently relies on a limited set of satellite observations. We have included a discussion of potential future enhancements, such as incorporating additional satellite data and integrating conventional observations, in the revised Discussion section (lines 455–462) of the main text.

Considered in terms of the *end-to-end* system, through joint optimization of analyses and forecasts *FuXi Weather* is able to produce forecasts that are competitive with or (in certain regions/variables/lead-times) better than HRES, using only satellite data in the inference stage. We hope this helps clarify the extent to which this is a significant milestone for ML systems. Revised sentences lines 192-196 are intended to better set expectations on this. At the same time, revisions in section 3.3 elaborate on specifics and caveats.

Reference:

Allen, A., Markou, S., Tebbutt, W. et al. End-to-end data-driven weather prediction. Nature (2025). https://doi.org/10.1038/s41586-025-08897-0

Further, when the Authors compare forecast results from FuXi Weather with those from ECMWF IFS, they are mixing the effects of two separate components, the assimilation system and the forecast model. If they want to evaluate how good the assimilation system of FuXi Weather is, then they need to use the same forecast model (in this case FuXi) started from a separate set of initial conditions (eg, ERA5) as a control. Otherwise the different characteristics of the forecast models used in the comparison muddy the conclusions. This is especially the case with MLWP models like FuXi which

are known to suffer from increasing smoothness with forecast lead time and thus can game standard performance metrics.

More detailed comments are provided in the attached annotated version of the manuscript.

From my perspective, this is a potentially interesting but very preliminary investigation and certainly at too early a stage to be considered for publication.

Response:

We appreciate the reviewer's comment on the importance of isolating the performance of the DA system from that of the forecast model. However, the primary objective of this study is to introduce FuXi Weather, a data-to-forecast system integrating DA and forecasting, rather than focusing solely on the DA component.

In Section 9 of the Supplementary Information, we have presented a targeted evaluation of four training configurations, including one (Setting 3 in Supplementary Table 1 and Supplementary Figure 24) that uses the original FuXi forecast model (without fine-tuning) and is initialized using FuXi-DA analyses. This directly responds to the reviewer's request for a controlled comparison to isolate the DA component's contribution.

We also agree with the reviewer that overly smooth forecasts remain a common limitation in machine learning-based weather forecasting models. To improve the clarity and comprehensiveness of the manuscript, we have made the following revisions:

- We have included a discussion in the main text (lines 375-381 and 385-387) to carefully caveat the role of reduced forecast activity at long lead times in enhancing forecast performance.
- To quantitatively assess forecast smoothness, We have introduced additional evaluation metrics in Section 4 of the Supplementary Information, including: 1) the standard deviation of forecast anomalies relative to climatological means (Bouallègue et al. 2024), and 2) the RMSE between forecasts and climatological means (Allen et al., 2025).
- Comparative results on forecast activity are presented in Section 11 of the Supplementary Information. We also analyze the trade-off between forecast accuracy and activity, demonstrating that it is possible to enhance forecast activity without significantly degrading accuracy. This topic remains an active area of research, which we intend to explore further in future work.
- We have applied statistical significance testing throughout our results to strengthen the reliability of our findings.

- We have included a comparison between FuXi Weather forecasts and its analyses in the main text, illustrating that the initial errors are small and comparable to those of ECMWF HRES.

We do hope the reviewer will also take into account the diagnostics presented in Section 3.4, which examines the physical consistency of analysis changes associated with specific observations. Since ML DA does not inherently encode knowledge of atmospheric processes, we advocate this type of careful cross-checks for ensuring physical consistency. While this approach is standard in traditional DA, it remains novel in the ML literature. While we recognize the limitations highlighted by the reviewer, we believe this work represents a meaningful step forward in the advancement of fully data-driven weather forecasting systems.

Reference:

Allen, A., Markou, S., Tebbutt, W. et al. End-to-end data-driven weather prediction. Nature (2025). https://doi.org/10.1038/s41586-025-08897-0

Ben Bouallègue, Z., and Coauthors, 2024: The Rise of Data-Driven Weather Forecasting: A First Statistical Assessment of Machine Learning–Based Weather Forecasts in an Operational-Like Context. Bull. Amer. Meteor. Soc., 105, E864–E883, https://doi.org/10.1175/BAMS-D-23-0162.1

1. *Line 44 in the main text. Maybe "local disparities in forecast accuracies" otherwise this sentence does not make sense*

Response: As suggested by the reviewer, we have revised the wording in the abstract to "*local disparities in forecast accuracies*".

2. *Line 134-135 in the main text. "FuXi Weather integrates a substantially enhanced version of FuXi-DA [42] with fine-tuned FuXi." Does this mean that the DA system is cycled based on a model trained on ECMWF ERA5 reanalysis fields?*

Response:

We confirm that the DA system is cycled using FuXi-DA, which was trained with ECMWF ERA5 reanalysis data as the reference. To improve clarity, we have added an additional sentence in the main text (lines 138-139).

3. *Line 185-187 in the main text. "End-to-end training of FuXi Weather optimizes both the analysis and forecasts using ECMWF ERA5 reanalysis data at 0.25° resolution as the reference." This implies that FuXi Weather is still completely dependent on ECMWF ERA5 reanalysis data to learn both the DA update and the forecast step, right?*

Response:

We confirm that the training of FuXi Weather relies on ECMWF ERA5 reanalysis data for both the DA and forecasting components during the model development. To clarify this dependency, we have revised the main text accordingly: lines 138-139 in Section 1, lines 194-196 in Section 2, and lines 472-473 in Section 4.

4. *Line 217-218 in the main text. "This subsection evaluates the analysis fields of FuXi Weather against 42-hour FuXi forecasts initialized with ERA5, using ERA5 as the benchmark." I do not understand what this means. One usually evaluates forecast against verifying analyses, not the other way round. Are the Authors saying that they evaluate the quality of the FuXi Weather analyses against FuXi forecasts initialised by ERA5 analyses valid 42 hours earlier? If this is the case, then I do not see the point of the exercise.*

Response:

We thank the reviewer for highlighting this point, which helped us to clarify a previously ambiguous statement in the manuscript. We confirm that both the FuXi Weather analyses and the 42-hour FuXi forecasts (initialized with ERA5) are evaluated against ERA5, which serves as the reference. To improve clarity, we have revised the relevant sentence in lines 229-230 in Section 3.1 of the main text.

Additionally, we have included evaluations of the FuXi Weather forecasts against its analyses in Section 3.2 and 3.3 of the main text. Further details are provided in response to the reviewer's subsequent comments.

5. *Line 218-221 in the main text. "Extended Data Fig.2 presents the globally-averaged, latitude-weighted RMSE for two FuXi Weather configurations: one incorporating background forecasts and one without." My reading of Extended Data Fig.2 is that the quality of the FuXi Weather analysis fields with/without background is for most variables/levels inferior to that of a FuXi forecast started from ERA5 analyses valid 42 hours earlier. If this interpretation is correct, the obvious conclusion is that FuXi Weather analyses are very poor.*

Response:

We appreciate the reviewer's careful examination of Extended Data Fig. 2 and would like to provide additional context for interpreting these results.

- The comparison between FuXi Weather's analyses and the 42-hour FuXi forecasts (initialized from ERA5) should be considered in the context of the observational data available for assimilation. Although FuXi is trained using ERA5 as the reference, it currently has access to a subset of the observations included in the full ERA5 reanalysis, particularly lacking the extensive set of conventional observations. Moreover, FuXi Weather has been trained on just one year of data.

Therefore, it is not surprising that there remains nontrivial analysis error. Although we expect to improve on this in future, the current analyses are sufficiently accurate to yield slowly growing forecast errors, allowing FuXi's medium-range forecasts to remain competitive.

- To better illustrate this, we have added an additional curve to Figures 2 and 3 in the main text, showing initial error growth relative to FuXi Weather's own analysis. Even with imperfect analyses, we would argue that FuXi Weather already demonstrates promising performance for medium-range forecasting.
- We have carefully presented the relevant figures in a way that caveats the initial errors in the current analyses. However, we recognize that we may have accidentally overstated these limitations by not including a comparable illustration of the differences between ECMWF HRES analyses and ERA5. To address this, we have added two figures to the Supplementary Information: Supplementary Figure 16 shows that the FuXi analyses have activity comparable to ECMWF HRES, and Supplementary Figure 17 shows that the MBE relative to ERA5 is of similar magnitude between FuXi and ECMWF HRES analyses.
- To our knowledge, the only other system pursuing a similar approach is that of Allen et al. (2025) which produces analysis fields with RMSE values comparable to those of 2-day ECMWF HRES forecasts. This comparison is directly relevant to our results in Extended Data Fig. 2. In our study, FuXi Weather's analyses outperform the 42-hour forecasts (initialized with ERA5) for relative humidity at 300 and 500 hPa, and demonstrate comparable accuracy for temperature, geopotential, and wind at these levels. Some degradation is observed at 850 hPa, which we acknowledge as an area for improvement through the assimilation of additional observations. Several differences between the Aardvark and FuXi Weather systems are discussed in Section 7 of the Supplementary Information.

While we recognize that there remains room for further improvement, such as expanding the observational dataset and the training period, we believe that FuXi Weather's analyses achieve state-of-the-art performance in this rapidly evolving field. Therefore, we respectfully disagree with the characterization of these results as "very poor".

Reference:

Allen, A., Markou, S., Tebbutt, W. et al. End-to-end data-driven weather prediction. Nature (2025). https://doi.org/10.1038/s41586-025-08897-0

6. *Line 263-265 in the main text. "Although FuXi Weather initially had higher RMSE values than those of ECMWF HRES, it outperformed ECMWF HRES after a variable-specific lead time, which varied according to the variable and pressure level."*

Line 273-276 in the main text. "However, Fuxi Weather forecasts initialized with analysis incorporating background forecasts, though initially less accurate than ECMWF HRES, improved over time and eventually achieved higher ACC values across all examined variables."

The only way I can think of that a forecast model having worse performance in the first 5-6 days than a control then becomes better is by losing forecast activity, i.e. becoming smoother with increasing forecast lead time. The Authors should provide forecast activity plots for all the variable shown in Figure 2 if these results are to be taken at face value.

Response:

We thank the reviewer for highlighting the importance of forecast activity analysis, which provides valuable context for interpreting our results. We acknowledge that our original comparison between FuXi Weather and ECMWF HRES requires further clarification.

The apparent initial RMSE advantage of ECMWF HRES arises because its forecasts are evaluated against its own analyses, as noted in the caption of Figure 2. This gives ECMWF HRES a favorable starting point, as its forecasts start with zero error relative to its own analysis fields. To provide a more equitable comparison, we have now included additional results in Figure 2, where FuXi forecasts are also evaluated against its own analyses. This adjustment allows for a more equitable assessment of the initial error growth in both systems, with each forecast verified against its respective analysis.

Regarding the reviewer's comment on forecast activity, we agree that this is an excellent point. Indeed, the relative performance improvement of FuXi at longer lead times may be partially attributed to reduced forecast activity, a phenomenon commonly observed in machine learning based weather prediction models. Additionally, FuXi demonstrates a stronger ability to reduce both systematic biases and random errors, as evidenced by the comparisons of MBE and standard deviations of error (STD_{ERROR}) in Extended Data Figures 5 and 8.

To quantify the degree of reduced activity in FuXi, we have conducted a comprehensive forecast activity analysis for all variables shown in Figure 2, using both the std-based method (Bouallègue et al., 2024) and the RMSE-based approach (Allen et al., 2025). Forecast activity was normalized relative to ECMWF HRES, with a value of 1 indicating activity equivalent to ECMWF HRES and values approaching 0 signifying convergence toward climatological means. These detailed results are now presented in Section 11 of the Supplementary Information, where we discuss the trade-off between forecast accuracy and forecast activity. Both the std-based and RMSE-based activity show similar trends when normalized against ECMWF HRES, providing essential context for understanding the performance characteristics discussed in the main results.

We have revised lines 376-393 in Section 3 of the main text to provide a more refined discussion of the forecast improvements in light of this trade-off, particularly during the interval before the loss of activity becomes substantial.We have also revised the

paragraph of the Discussion section (lines 463-467) to include caveats regarding this effect.

7. *Fig. 2 caption. "Comparison of forecast performance among various models over a 1-year testing period, spanning from July 03, 2023, to June 30, 2024." There is a notable and unphysical jump at forecast day 4 in the FuXi forecasts. Can the Authors explain where this originates?*

Response:

We appreciate the reviewer's observation regarding the discontinuity in forecast performance at day 4. This feature arises from the design of our FuXi Weather system, which employs a cascade model architecture, as described in our previous publication (https://www.nature.com/articles/s41612-023-00512-1). In that paper, FuXi consists of three cascade models, FuXi-Short (0-5 days), FuXi-Medium (5-10 days), and FuXi-Long (10-15 days), each optimized for a specific forecast window. As demonstrated in the Supplementary Information of this paper, 4-day forecasts generated by FuXi-Short when initialized with FuXi-DA analysis fields, achieve accuracy comparable to 5-day forecasts initialized with ERA5. Therefore, we implemented a configuration where FuXi-Short generates 0-4 day forecasts, which are then used as initial conditions for FuXi-Medium to produce forecasts for days 4-10. The discontinuity observed at day 4 reflects model transition.

To enhance methodological clarity, we have made two revisions to the main text. In Section 2 of the main text, we have revised the description in lines 189-191, and in Section 3, we have added clarifications in lines 295-297.

Additionally, we have updated the captions of Figure 2, Extended Data Figure 3, and Supplementary Figures 21 and 22 (now renumbered as Supplementary Figures 24 and 25) to include the following sentence: "The performance change at day 4 arises from the model transition from FuXi-Short to FuXi-Medium."

While this cascaded architecture introduces a discontinuity, it allows optimized performance across different forecast windows. We also appreciate this opportunity to clarify the methodology.

8. *Line 325-328 in the main text. "These findings suggest that FuXi Weather can deliver more accurate forecasts using an substantially less data than traditional NWP systems, offering a cost-effective, satellite-based solution to improve forecasts in regions lacking observational infrastructure." One can come up with a different explanation for the results shown over Central Africa. Central Africa is in the Tropics so the evolution of forecast errors is dominated by systematic errors and conditional biases. While the FuXi forecast starts from a clearly less accurate initial state, the fact that FuXi has been trained on ERA5 reanalyses means that it has*

learned to compensate for the IFS model biases. This, together with the reduced forecast activity of the FuXi model, can explain the results shown here.

Response:

We appreciate the reviewer's insightful analysis regarding the potential factors contributing to FuXi Weather's superior performance in central Africa. The suggestions about systematic error compensation and reduced forecast activity offer valuable alternative interpretations that merit careful consideration. To rigorously examine these hypotheses, we conducted three complementary analyses.

First, we have decomposed the forecast errors into systematic and random components by calculating the mean bias error (MBE) and the standard deviation (std) of errors (STD_{ERROR}). As illustrated in Extended Data Fig.5, FuXi Weather exhibits both lower MBE and smaller STD_{ERROR} values compared to ECMWF HRES, indicating more effective mitigation of both systematic biases and random errors. For clarity, we have revised lines 338-345 of the main text to explicitly state these findings.

Second, as presented in the updated Figure 3 and Extended Data Figure 5, we have expanded the evaluation of FuXi Weather forecasts against its own analyses over central Africa. The results demonstrate that FuXi Weather (green lines) consistently outperforms ECMWF HRES (blue lines) throughout the entire 10-day forecast period when each is verified against its respective analyses.

Lastly, we calculated forecast activity, defined as the std of forecast anomalies relative to climatological means, over central Africa. The results reveal that FuXi Weather maintains lower forecast activity than ECMWF HRES from the start of the 10-day forecast period, suggesting smoother predictions. This is a typical characteristic of machine learning-based weather prediction models. Notably, FuXi Weather's superior skill (red lines) becomes apparent prior to any considerable reduction in forecast activity. This implies that while reduced activity may favor performance at longer lead times, it does not fully account for FuXi Weather's enhanced accuracy beyond the initial forecast period. We have revised the lines 350-359 to improve clarity on this point. Additionally, as detailed in our response to the reviewer's Comment #6, we have explored the trade-offs between forecast accuracy and activity, including preliminary results on an enhancement module designed to improve forecast activity.

We have updated the relevant figures and corresponding text to reflect these comprehensive analyses. While we acknowledge that the factors highlighted by the reviewer may contribute partially, our extended analyses support the original conclusion that FuXi Weather demonstrates robust and competitive performance in data-sparse regions.

Reply to Reviewer #3

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #3 (Remarks to the Author):

The authors present an evolution of their machine-learning (ML) based forecasting system emulator Fuxi that is assimilating selected satellite based observations to create initial conditions for forecasts instead of using external analyses for initialisation. The data assimilation replacement method, called Fuxi-DA, is combined with an enhanced Fuxi forecast emulator to create the end-to-end Fuxi-Weather system that can be cycled like traditional, numerical methods based analysis-forecast systems. The output is assessed using standard forecast error metrics, and a specific focus is applied to a region in central Africa with sparse observational data coverage to test the robustness of the method.
The paper is complete regarding the explanation of methods, data, experiments and evaluation.
The Fuxi developers build on an extensive record of ML applied to weather forecasting,and they exhibit deep knowledge of the methodological frameworks applied in the numerical weather prediction (NWP) community. They have also published their results before and made code available for others to try - which is highly commendable. The paper is very well written.
In the attached file, I have raised several points that will need addressing before publication.
My greatest concern is the lack of evidence of statistical significance of forecast performance when compared against other data. More information is needed on the use of observational data in the data-assimilation process. I also wonder about the future need of data assimilation as an intermediate step between observational data records and forecast generation. Lastly, while I appreciate a special focus on central Africa for the reasons explained in the paper, the evaluation of performance for this case is not overly convincing.

Summary:

The authors present an evolution of their machine-learning (ML) based forecasting system emulator Fuxi that is assimilating selected satellite based observations to create initial conditions for forecasts instead of using external analyses for initialisation. The data assimilation replacement method, called Fuxi-DA, is combined with an enhanced Fuxi forecast emulator to create the end-to-end Fuxi-Weather system that can be cycled like traditional, numerical methods based analysis-forecast systems. The output is assessed using standard forecast error metrics, and a specific focus is applied to a region in central Africa with sparse observational data coverage to test the robustness of the method.

The paper is complete regarding the explanation of methods, data, experiments and evaluation. The Fuxi developers build on an extensive record of ML applied to weather forecasting, and they exhibit deep knowledge of the methodological frameworks applied in the numerical weather prediction (NWP) community. They have also published their results before and made code available for others to try - which is highly commendable. The paper is very well written.

In the following, I have raised several points that will need addressing before publication. My greatest concern is the lack of evidence of statistical significance of forecast performance when compared against other data. More information is needed on the use of observational data in the data-assimilation process. I also wonder about the future need of data assimilation as an intermediate step between observational data records and forecast generation. Lastly, while I appreciate a special focus on central Africa for the reasons explained in the paper, the evaluation of performance for this case is not overly convincing.

General comments:

•1 Need for data assimilation: Given the recent evolution of ML-methods applied in weather prediction, I wonder whether there will still be a need for an explicit, intermediate data assimilation step in the future. The present need for this step originates from numerical models requiring starting conditions to evolve the differential equations in time. ML methods do not necessarily need this unless they cycle through time steps like traditional systems. Recent studies seem to suggest that it may be advantageous to produce predictions directly from past and present observations, and therefore skip the data assimilation step all together (aka direct observation prediction, e.g. https://arxiv.org/abs/2407.15586).

While this point does not negate the importance of assessing ML-methods for data assimilation per-se, this paper could benefit from a discussion of this topic given that the target journal is Nature Comm.

Response:

We appreciate the reviewer's insightful comment regarding the potential evolution of machine learning-based weather prediction systems and the possibility of eliminating explicit data assimilation (DA) steps. Indeed, recent pioneering efforts, such as Artificial Intelligence Direct Observation Prediction (McNally et al., 2024; *https://arxiv.org/abs/2407.15586*) and GraphDOP (Alexe et al., 2024; https://arxiv.org/html/2412.15687v1), have demonstrated the feasibility of generating weather forecasts directly from meteorological observations using machine learning. Developed at ECMWF, both approaches offer significant advantages by obviating traditional DA challenges, including the estimation of error covariance matrices, and by enabling the assimilation of all available observations regardless of measurement complexity.

However, the effectiveness of AI-DOP frameworks depends critically on two prerequisites: (1) sufficiently dense spatiotemporal observational coverage and (2) access to long-term, high-quality historical observational records. For example, ECMWF's AI-DOP model is trained on 18 years of observational data (2004–2021), covering all major categories used in numerical weather prediction (NWP) systems. In contrast, FuXi Weather has thus far been trained on only one year of data. Explicit DA approaches, which typically rely on forecasting models pretrained using multi-decade ERA5 reanalysis datasets, are better suited to scenarios with sparse or incomplete observational data. Implicit DA methods require significantly larger and more consistent observational datasets to learn direct observation to prediction directly.

Additionally, our objective requires generating forecasts that can be directly evaluated against ERA5 and compared with state-of-the-art NWP models, i.e., ECMWF HRES. Such a comparison must encompass not only surface parameters measured by weather stations but also upper-air variables like 500 hPa geopotential height, for which direct observations are unavailable.

Following the reviewer's suggestion, we have expanded the Discussion section (lines 473–500) to explicitly address this important consideration.

•2 Statistical significance: It is standard practice in NWP evaluation to indicate the statistical significance of skill (differences). If the individual variance of two skill measures is too large, it will not allow to make statements about the significance of their difference:"Uncertainty in the measures can be indicated as confidence intervals or error bars calculated using approximations to statistical distributions, or through computer-intensive techniques such as re-sampling"
(https://empslocal.ex.ac.uk/people/staff/dbs202/publications/booksreports/mailier.pdf).

This applies to all figures 2, 3 and extended figures 3, 4, 5, 8, also maps. The interpretation of these figures will have to be adjusted as well as several score differences may not lead 1 to conclusions on better or worse performance. From my experience, some of the stated score differences will become meaningless once information on statistical significance is added.

Response:

We appreciate the reviewer's valuable suggestion regarding the importance of statistical significance testing in NWP evaluation. We have conducted a comprehensive significance testing analysis for all relevant figures, following the established methodology described by Geer (2016). This includes both the time series of forecast performance (Figures 2 and 3; Extended Data Figures 2, 3, 5, and 8; and Supplementary Figures 14 and 15, previously numbered 13 and 14) and spatial distributions of forecast errors (Supplementary Figures 18-23, previously numbered 15-20).

In the time series plots, statistically significant differences are now indicated by solid dots overlaid on the corresponding lines. In the spatial error maps, stippling has been added to denote regions where differences in skill scores are statistically significant. We have revised the relevant sections of the manuscript to clearly reflect these updates.

The inclusion of statistical significance strengthens our interpretations of the results without changing the study's main conclusions. The key differences in model performance remain statistically significant, thereby increasing confidence in the robustness of our evaluation.

Reference:

Geer, A.J.: Significance of changes in medium-range forecast scores. Tellus A: Dynamic Meteorology and Oceanography 68(1), 30229 (2016). https: //doi.org/10.3402/tellusa.v68.30229

•3 Observation pre-processing: The satellite radiance observations are 'assimilated' in a rather unqualified way. All channels of the microwave sounding instruments are used even though the window channels of both temperature and humidity sounders are very sensitive to surface emission that is driven by emissivity (dependent on surface type, moisture, vegetation, roughness) and surface temperature. These channels are also sensitive to clouds and precipitation, which greatly dilute the information content on the five target state variables. The authors mention cross-track scanning instrument (scan angle) biases. Are these corrected or is it assumed that - through knowledge of the scan-angle -the ML-method adjusts its weights? Can this be shown?

The GNSS radio occultation data is used as profiles of refractive index, and this product exhibits larger errors in the lower 2-3 km of the troposphere and in the upper stratosphere/mesosphere. Pre-processing is mentioned, but the paper does not provide information on how systematic errors associated with this data are being addressed.

Overall, the study uses observational data in a brute-force fashion that could be greatly improved on given the available knowledge from traditional data assimilation.

Response:

We appreciate the reviewer's insightful comments about observation pre-processing. While we acknowledge that incorporating established DA knowledge into observation handling could potentially enhance model performance, our approach intentionally adopts a more direct strategy. This design allows the model to learn the relationships among inputs, such as satellite observations, scan angle and spatial coordinates, and the target meteorological variables during training.

To assess how effectively FuXi-DA captures these relationships, we conducted a gradient-based sensitivity analysis. Specifically, we computed the gradients of relative

humidity (R) with respect to five ATMS humidity channels in the 183.31 GHz band (channels 18-22), scan angle, latitude, and longitude. For R at each pressure level, we calculated a latitude-weighted MAE loss relative to ERA5. The resulting gradients preserve the input dimensions of the ATMS observations ($8 \times 8 \times 720 \times 1440$, where the second dimension of 8 includes five channels plus three additional dimensions for scan angle and spatial coordinates). To quantify the magnitude of sensitivity, we took the absolute values of these gradients and aggregated them across the temporal and spatial dimensions (8, 720, 1440), resulting in eight gradient sums (one for each ATMS channel), along with aggregated contributions from scan angle, latitude, and longitude. These values were then normalized to the [0, 1] range using min-max scaling, with the minimum and maximum determined across all variables and channels.

The resulting sensitivity matrix (13×8) reveals the relative influence of each ATMS channel, scan angle, and spatial coordinates on the analysis fields. These results confirm that FuXi-DA learns to weight input in a physically meaningful manner, even without explicit bias corrections.

We have also added a new Section 10 in Supplementary Information to provide further clarification on how the model accounts for potential biases and uncertainties in the input data.

•4 African case study: I understand the motivation of the regional focus when applied to an area that is important for society, sparsely observed and where traditional system do not perform well. However, the study does not spend enough effort to fully assess the African case. Fig. 3 shows performance, again without statistical significance information, a two-cycle/day skill-score pattern (?), forecasts are verified against analyses that are notoriously poor in this area as opposed to (the few existing) observations, and for parameters that matter less for society here than, for example, soil moisture and rainfall.

This is not a very convincing.

Response:

We thank the reviewer for the valuable feedback regarding the African case study. We have conducted additional analyses to further characterize FuXi Weather's performance over central Africa.

First, we have updated Figure 3 to include statistical significance testing, where red dots indicate time steps at which FuXi Weather shows statistically significant improvements over ECMWF HRES. As shown in the updated Figure 3 and Extended Data Figure 5, we have expanded the evaluation of FuXi Weather forecasts by comparing them against its own analyses. The results indicate that FuXi Weather (green lines) outperforms ECMWF HRES (blue lines) in forecasting U850, T2M, and MSLP across the 10-day forecast period.

To include variables of greater societal relevance, we extended the evaluation to total precipitation (TP), as FuXi Weather does not output soil moisture. The updated Extended Data Figure 5 presents TP forecast performance comparisons over central Africa. Since ECMWF HRES-fc0 contains zero precipitation values, both models are evaluated against ERA5 TP. Results suggest that FuXi Weather achieves better TP accuracy, particularly for light precipitation events, as further illustrated in Extended Data Figure 6.

Additionally, we have analyzed forecast activity, defined as the standard deviation of forecast anomalies relative to climatological means. This analysis reveals that FuXi Weather maintains lower activity than ECMWF HRES over the 10-day forecast period, suggesting smoother predictions. This is a characteristic commonly observed in machine learning-based weather prediction models. As shown in Extended Data Figure 5, FuXi Weather's enhanced skill (red lines) becomes apparent before considerable reduction in forecast activity. This suggests that reduced forecast activity may partially contribute to FuXi Weather's superior performance at longer lead times. Furthermore, FuXi exhibits a greater capacity to reduce both systematic biases and random errors, as evidenced by the comparisons of MBE and standard deviations of error (STD_{ERROR}) in Extended Data Figures 5 and 8.

We have also explored the trade-offs between forecast accuracy and activity, including preliminary results demonstrating the efficacy of an enhancement module designed to improve forecast activity.

We acknowledge the reviewer's valid concern about verification against ERA5 reanalysis data or ECMWF HRES analyses, both of which assimilate limited observations in central Africa. Therefore, we have revised Section 3.3 to present more cautious interpretations of our results and explicitly acknowledge the limitations of our evaluation. Specifically:

1) Lines 326–336 have been revised to describe the results of the expanded evaluation.
2) We have removed the previously overstated claim: "*These findings suggest that FuXi Weather can deliver more accurate forecasts using substantially less data than traditional NWP systems, offering a cost-effective, satellite-based solution to improve forecasts in regions lacking observational infrastructure.*"
3) We have added a discussion of evaluation limitations in lines 383-393, emphasizing that while preliminary results indicate FuXi Weather may produce forecasts of comparable or potentially superior accuracy using fewer observations, further validation with independent in-situ datasets is essential.

These revisions provide a more balanced interpretation. We agree that future work would benefit from validation using independent observational datasets, and have toned down our conclusions accordingly.

Specific comments:

• L56-61/L152-154, and throughout: See General Comments on performance outcomes given statistical significance above, and from selected comments below.

Response:

As suggested by the reviewer, we have conducted significance testing for all relevant figures and revised corresponding text in the manuscript.

• L89: Please also mention efforts for ensemble forecasts towards uncertainty quantification for both traditional (which also feed off ensemble analyses) and ML methods.

Response:

Following the reviewer's suggestions, we have added sentences in lines 87-89 of the Introduction to describe recent efforts toward developing ensemble forecasts using machine learning methods.

• L207-212: Fuxi forecasts are evaluated with ERA-5, but ERA-5 is based on ECMWF's operational system from 2016 (nearly 10 years ago!), it has fairly coarse resolution (nominally 31 km, but analysis increments are produced at 124 km!).

HRES seems to be evaluated with HRES at t0. This will favour HRES, and also affect the regional assessment over Africa with strong orographic variability that ERA-5 will not resolve (, and for which MSL is not the best parameter to choose by the way). RMSE evaluations may actually suffer from higher resolution as increased variability penalizes RMSE.

Further, the NWP community does indeed evaluate forecasts with both own and reference analyses. Ideally, you show both and include HRES vs ERA-5.

Response:

As suggested by the reviewer, we have expanded the evaluation of FuXi Weather forecasts by evaluating them against its own analyses, which are represented by green lines in Figures 2 and 3, as well as in Extended Data Figures 3, 5, and 8.

• Fig. 2 and others: I assume the dip in forecast skill for Fuxi at day 4 is explained by the switch from Fuxi-Short to Fuxi-Medium. This information could be added in the caption.

Response:

Following the reviewer's suggestion, we have added the following sentence to the captions of Figure 2, Extended Data Figure 3, and Supplementary Figures 24 and 25 (previously numbered 21 and 22):

"*The performance change at day 4 arises from the model transition from FuXi-Short to FuXi-Medium*."

Additionally, we have added a sentence to the captions of Supplementary Figures 6 and 7 (previously numbered 5 and 6).

"*The performance change at days 4 and 5 reflects the transition from FuXi-Short to FuXi-Medium for the respective forecasts (red and gray lines)*."

• Section 3.1: Would it not be informative to compare global analyses fields to begin with, like Fuxi-DA vs ERA5 vs HRES at t0 for MSL, t2m and u/v300? This will exhibit, for example, the analysis activity = standard deviation / climatological mean, which is important for RMSE evaluations, but also mean differences, for example, systematically colder/warmer or windier/calmer analyses.

Response:

As suggested by the reviewer, we have conducted comprehensive comparisons of analysis activity and mean bias error (MBE) between FuXi Weather and ECMWF HRES, as presented in the new Section 7 of the Supplementary Information. The mathematical formulation for analysis activity is now provided in Section 4 of the Supplementary Information.

• Line 286-288: Global model performance is actually evaluated for regions, for example, Europe, North America, East Asia, Australia etc. While this is not tied to socioeconomic implications, it is still a regional evaluation.

Response:

Following the reviewer's suggestions, we have revised the sentence in lines 313-316 of the main text.

• L308-310: So where does the enhanced skill of Fuxi come from if ERA-5 is troubled over less well observed regions, surface in-situ observations are missing, and satellite data is affected by all kinds of signals from lower atmosphere, clouds, precipitation and surface characteristics (see above)?

Response:

We appreciate the reviewer's insightful question regarding the superior forecast skill of FuXi Weather. As elaborated in our response to the reviewer's comment #4 about the African case study, we have conducted additional analyses to investigate why FuXi outperforms ECMWF HRES, particularly in regions with sparse surface in-situ observations.

First, we have decomposed forecast errors into systematic and random components by computing the mean bias error (MBE) and the standard deviation (std) of errors (STD_{ERROR}). Extended Data Fig.5 demonstrates that FuXi Weather exhibits both lower MBE and smaller STD_{ERROR} values than ECMWF HRES, indicating its improved ability to reduce systematic biases and random errors.

Second, we have analyzed forecast activity, defined as the std of forecast anomalies relative to climatological means. The results show that FuXi maintains lower forecast activity than ECMWF HRES throughout the 10-day forecast period, suggesting smoother predictions. This is a characteristic often associated with machine learning-based weather predictions. Notably, FuXi Weather's superior skill (red lines) becomes apparent prior to any considerable reduction in forecast activity.

In summary, FuXi's improved performance compared to ECMWF HRES is likely attributable to two key factors: its enhanced ability to reduce both systematic biases and random errors, and its lower forecast activity, which together contribute to more accurate and stable forecasts. For clarity, we have revised lines 338-359 of the main text to explicitly state these findings.

• L312: 'performed or 'outperformed'?

Response:

Thanks for highlighting this typo, the text has been corrected by changing "*performed*" to "*outperformed*".

• Section 3.4: I like the use of single-observation experiments. For clarity: are you showing increments or increment differences between the analyses from using a perturbed minus an unperturbed ATMS radiance? It would be great to see whether a 3D-Var would show the same increments or their differences.

Response:

We thank the reviewer for the valuable suggestions regarding Section 3.4. We acknowledge that the original use of the term "*analysis increment*" was inaccurate, as this section actually examines differences between analysis fields, derived from perturbed and unperturbed ATMS radiance observations. To improve clarity, we have carefully revised the terminology throughout Section 3.4 accordingly. Specifically, we have changed the title of Section 3.4 to "*Physical consistency of* ***analysis changes***".

Related revisions have been made to several sentences within Section 3.4 of the main text (lines 398, 402, 411–413, and 426), as well as in the caption of Figure 4. Corresponding revisions have also been made in Section 5.1 of the Supplementary Information (lines 511 and 539).

Regarding the reviewer's suggestion to compare our results with a 3D-Var system, we appreciate this thoughtful recommendation. However, such a comparison is technically challenging, as FuXi-DA outputs atmospheric profiles at only 13 pressure levels. This relatively coarse vertical resolution is inadequate for computing accurate Jacobian matrices required by fast radiative transfer models typically employed in 3D-Var data assimilation systems. While this limitation prevents us from conducting the comparison in the present study, we recognize its potential value and plan to explore higher vertical resolution in future work to facilitate such analyses.

• Supplement L190-193: If not treated properly (see General Comments item no. 3), the error characteristics from different instruments on different satellites become merged when averaging to the Fuxi grid!

Response:

We appreciate the reviewer's valuable suggestions. As addressed in our response to the reviewer's Comment #3, we acknowledge that incorporating established domain knowledge from traditional data assimilation frameworks, particularly by accounting for the distinct error characteristics of different satellite instruments, could further improve model performance. We plan to pursue this direction in future work.

• Supplement L211.1: I am not sure I fully understand how the PointPillars method is applied to GNSS RO data. Can the 3d-aspect of this be explained with a figure?

Response:

We appreciate the reviewer's suggestion and have added Supplementary Figure 4 to illustrate the application of the PointPillars method in processing GNSS-RO data.

Additionally, we have revised the corresponding text in lines 212-245 of Section 2.1 in the Supplementary Information to improve clarity and readability.

• Supplement Section 5: This section seems too long, and 3D-Var/Jacobians are standard knowledge. I suggest to reduce this section and focus on the single-observation experiment set-up, a summary of the results and in how far they deviate from, say, 3D-Var.

Response:

We acknowledge the interdisciplinary nature of this study and recognize the importance of ensuring readability for readers who may be less familiar with specific concepts. Therefore, we have retained the relevant explanatory content and figures.

2nd Review of "FuXi Weather: A data-to-forecast machine learning system for global weather"

Reply to Reviewer #1

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #1 (Remarks to the Author):

The authors did a lot of work clarifying parts of the manuscript based on reviewer suggestions, which I appreciate. I have a few remaining comments.

Comments

1. *Introduction and conclusion: It could be worth emphasizing that this is the first ML system that can cycle in a DA-like system for weather forecasting (to my knowledge, at least.) Specifically, this is the only system that I know of that can produce forecast fields that feed back into the system without blowing up. This, I think, is a major achievement. For example, Slivinski et al (2025) found that most deterministic ML models blow up within a few weeks of cycling in a traditional DA system.*

Response:

We appreciate the reviewer's suggestion and have explicitly emphasized that FuXi Weather is the first machine learning system capable of performing both cycling data assimilation and forecasting for weather prediction. The "**Abstract**" (line 56), "**Introduction**" (lines 161-163) and "**Discussion**" (lines 468-469) sections have been revised accordingly.

2. *Lines 160-161 "Furthermore, FuXi Weather's development costs are considerably lower than traditional NWP systems." This is not strictly true. FuXi Weather's development relies heavily on ERA5, which built upon decades of NWP and reanalysis development. This is an important technicality, since many ML papers tend to gloss over this fact. With the exception of the AI-DOP techniques that learn directly from observations without reanalyses, every ML weather model is trained on a dataset that took years to develop and produce, and was built upon decades of knowledge.*

Response:

We appreciate the reviewer's insightful comment regarding the development costs of ML forecasting systems such as FuXi Weather. We have removed the original sentence in **"Introduction"** section and added new statements in line 502 of the "**Discussion**" section to acknowledge that ERA5 took several decades to develop.

3. *Line 208: "a 1-year data" should be maybe "one year of data"*

Response:

Thanks for highlighting this issue, we have fixed it.

4. *Lines 213-214: Is the variant of FuXi-DA without background forecasts discussed in the supplementary information? If not, I would consider adding it. Is this essentially direct-from-observation prediction?*

Response:

We confirm that the variant of FuXi-DA without background forecasts is essentially direct-from-observation prediction, and we have added this description in lines 219-220 in Section 3 of the main text. Following the reviewer's suggestion, we have also included a brief discussion on the importance of background forecasts in data-to-forecast system in the newly added Section 9 of the Supplementary Information.

5. *Extended Data Fig 2: maybe replace "FuXi analyses" with "FuXi Weather analyses" to make the distinction between the two systems more clear?*

Response:

Following the reviewer's suggestion, we have revised aption of Extended Data Fig 2 by replacing "FuXi analyses" with "FuXi Weather analyses".

6. *Extended data Figs 5-6: I see that the authors added precipitation per a reviewer comment. However, there are some major issues with precipitation in ERA5 especially in the Tropics (Lavers et al 2022), such that this comparison is not very meaningful. I would urge the authors to consider the reviewer's other suggestion, to compare to observations; even a gridded observational product like GPCP (https://climatedataguide.ucar.edu/climate-data/gpcp-monthly-global-precipitation-climatology-project) or IMERG (https://gpm.nasa.gov/data/imerg) would help provide context here.*

Response:

As suggested by the reviewer, we have added Extended Data Fig.9, which evaluates precipitation forecasts using IMERG instead of ERA5. The results yield slightly

different conclusions: while FuXi Weather continues to outperform ECMWF HRES in terms of RMSE, it shows lower ACC and higher absolute MBE.
These findings are discussed in the revised Section 3.3 (lines 391-401), where we discuss the precipitation forecast performance comparisons using IMERG as the reference dataset. The references suggested by the reviewer also been added.
We have also added lines 505-510 in "**Discussion**" section to highlight the potential for improved FuXi precipitation forecasts when trained and evaluated against more accurate observational datasets such as IMERG.
Additionally, the "**Data Availability Statement**" section (lines 548-459) has been updated to include information on accessing the IMERG dataset.

7. *I still don't understand the reasoning behind using overlapping DA windows. Why was this design choice made? Can you articulate the motivation behind this design choice in the paper? Also, I would recommend explicitly stating in the supplementary information that some observations are used more than once, if that is the case.*

Response:

Thank you for your valuable comment. We have now explicitly clarified the use of overlapping DA windows in Section 1.2, lines 94-101 of the Supplementary Information. The choice of an 8-hour window is based on practical considerations related to deep learning implementation. Specifically, the number 8 was chosen because it is a power of two, which aligns well with the computational architecture of GPUs. Using dimensions that are powers of two allows for more efficient memory usage and computational performance, due to the way parallel processing is handled by GPUs.

8. *Supplementary Info line 538: "a small perturbation at a single grid point." If I understand correctly, this is inaccurate, since the perturbation is in fact over several grid points, correct?*

Response:

Thanks for highlighting this issue, we removed the word "*at a single grid point*".

9. *Captions of supp figs 18-22 and related description: May want to clarify that these are differences of RMSEs, as opposed to RMS differences, and define the direction of the difference. It would be helpful to add the description of what red vs blue means in the figure captions as well.*

Response:

As suggested by the reviewer, we have revised the captions of the original Supplementary Figures 18–23 (renumbered as Supplementary Figures 15–20) accordingly.

References:

Slivinski, Laura C., et al. "Assimilating observed surface pressure into ML weather prediction models." Geophysical Research Letters 52.6 (2025): e2024GL114396.

Huffman, George J., et al. "Improving the global precipitation record: GPCP version 2.1." Geophysical Research Letters 36.17 (2009).

Kirschbaum, D.B.; Huffman, G.; Adler, R.F.; Braun, S.; Garrett, K.; Jones, E.; McNally, A.; Skofronick-Jackson, G.; Stocker, E.; Wu, H.; et al. NASA’s Remotely Sensed Precipitation: A Reservoir for Applications Users. Bull. Am. Meteorol. Soc. 2017, 98, 1169–1184.

Lavers, D.A., Simmons, A., Vamborg, F. & Rodwell, M.J. (2022) An evaluation of ERA5 precipitation for climate monitoring. Quarterly Journal of the Royal Meteorological Society, 148(748) 3124–3137. Available from: https://doi.org/10.1002/qj.4351

Reply to Reviewer #2

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #2 (Remarks to the Author):

I wish to thank the Authors for their detailed responses and the time and effort devoted to run the additional experiments and produce the new diagnostics. Unfortunately, I am still not convinced that the manuscript is ready for publication.

My main concern, among others, was that Fuxi-DA produces analyses that when used to initialise a forecast model generate forecasts that have ~2 days lower skill than those initialised with ERA5 initial conditions. To put things into perspective, a 2-day forecast skill drop is what we would see in a state of the art NWP DA system if all satellite observations were withheld. Bear in mind that satellite observations constitute ~95% of all assimilated observations in current operational NWP.

The Authors have two answers to this: 1) This is comparable to the state of the art in ML-driven DA, citing Allen et al. 2025 (https://www.nature.com/articles/s41586-025-08897-0); 2) The amount of observations assimilated by Fuxi-DA is smaller than that used in ERA5, especially considering the absence of conventional observations.

My perspective is that even though results of Allen et al., 2025, were not particularly convincing, results presented here do not show a significant improvement on those.

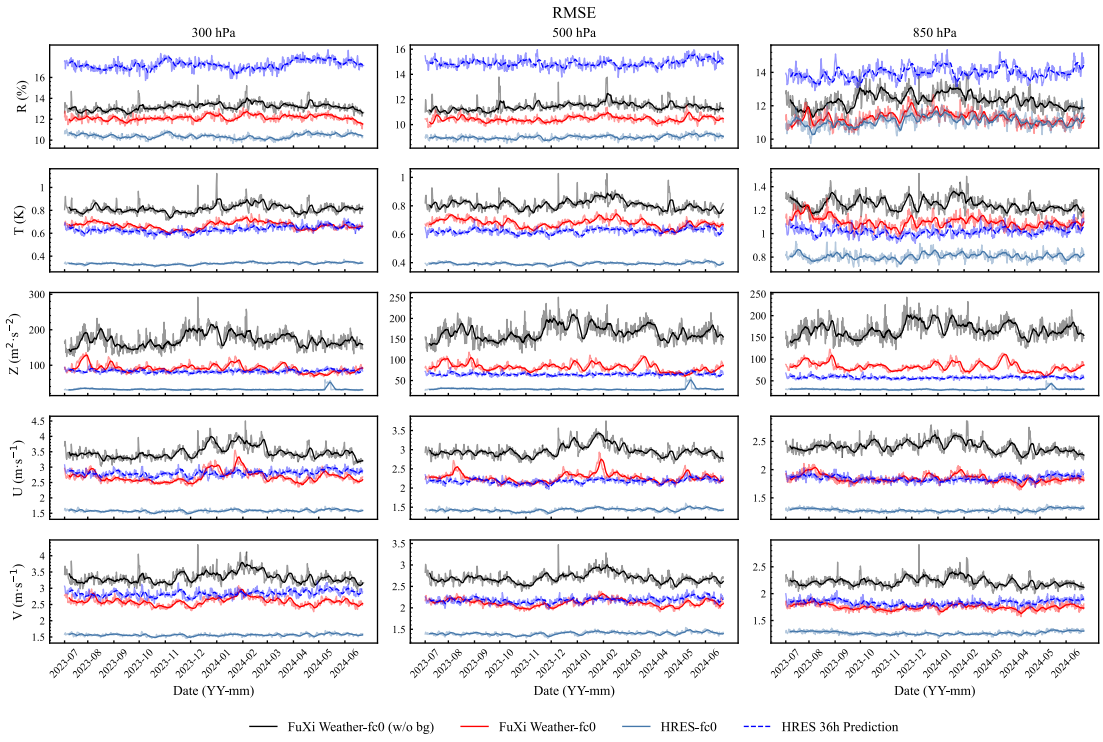
My suggestion to the Authors is to further develop their system by suitably extending the range of observations used, and only at that point it would be clearer whether end-to-end ML DA is a serious candidate for operational/reanalysis DA.

Response:

We sincerely appreciate the reviewer's thoughtful and constructive comments. To provide a broader perspective on the differences among various analyses, we have included, for the Reviewer's interest, the RMSE of HRES-fc0 and 36-hour ECMWF HRES forecasts evaluated against ERA5, using a format similar to that of Extended data Fig. 2. As expected, for U, V, Z, and T, FuXi Weather analyses do not perform as well as HRES-fc0, primarily due to the significantly smaller number of observations used. Specifically, FuXi Weather relies on data from five microwave instruments, whereas ECMWF HRES incorporates observations from approximately ninety satellite instruments. However, for relative humidity, particularly at 850 hPa, FuXi Weather analyses are nearly as accurate as the RMSE between HRES-fc0 and ERA5. Overall, FuXi Weather analyses achieve performance comparable to that of 36-hour ECMWF HRES forecasts, and for variables such as relative humidity, FuXi Weather analyses demonstrate notably higher accuracy.

We have not added this additional figure in the manuscript, as in the rest of the paper, ECMWF HRES forecasts are evaluated against HRES-fc0, and we do not feel that the 0 lead time analysis error is the primary point of our study. We certainly agree with the reviewer that reducing this error is a desirable goal for the DA component of the system. However, for this stage the early time error, while carefully documented, is not the primary focus. We hope we can convince the reviewer of the value of demonstrating an end-to-end system that yields competitive medium and long lead forecasts even without using all available data.

We believe this paper is already comprehensive, demonstrating a machine learning–based system capable of cycling data assimilation and forecasting, marking a significant development in its own right. We welcome the opportunity to include additional observations in future studies, where more extensive results can be presented and discussed relative to the current baseline.



Reply to Reviewer #3

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #3 (Remarks to the Author):

Summary:
The authors revised their paper in response to three reviews. The clarity of the paper has greatly improved but I still feel that it needs several corrections and changes before publication.

Comments:

• Abstract: ECMWF would claim that they are running an all-grid, all-sky/surface, allchannel forecasting system, which Fuxi is actually benefiting from because ERA-5 supplies the training data. ECMWF's exclude selected observations because of too large uncertainties due to cloud/precipitation/surface contamination. However, the fact that Fuxi includes all radiometer channels in their training (in a rather unqualified manner) does not mean that all channels contribute useful information in all situations. Fuxi ould have to mimic ECMWF's observation screening to understand whether the more aggressive approach is beneficial or not.

Response:

We appreciate the reviewer's comment regarding whether all satellite data from all channels provide useful information. We have removed the claim in the "**Abstract**" that FuXi is the first system to achieve all-grid, all-surface, all-channel, and all-sky DA.

• Section 3.2 and Fig. 3: Fuxi seems to outperform ECMWF mostly for days 6 and beyond for the important variables, namely geopotential height, temperature and wind. The more general statements on Fuxi outperforming ECMWF in the Abstract should therefore be more specific. I really appreciate the addition on analysis and forecast activity that was also requested by Reviewer no. 2. However, when looking at Fuxi's forecast activity, even after its enhancement, it is still substantially lower than that of ECMWF. This happens to coincide with the forecast range where Fuxi appears to be better. A sound conclusion could be that Fuxi only beats ECMWF forecasts at forecast ranges where it is quieter and will therefore produce better RMSEs.

The claims on performance in both Abstract and Discussion need to be very clear on this.

Response:

Following the reviewer's suggestion, we have revised the statement in the Abstract to clarify that "FuXi Weather consistently outperforms ECMWF HRES in observation-sparse regions beyond day one". Additionally, in the "**Discussion**" (lines 492-494), we have included a statement highlighting that FuXi outperforms ECMWF HRES in global forecasting performance at lead times where FuXi's forecasts exhibit greater smoothness.

• Section 3.3: I still think that verification against analyses in these areas is not very useful, and that a verification against observations would carry a lot more weight!

Response:

As suggested by the reviewer, we have added Extended Data Fig.9, which evaluates precipitation forecasts using IMERG instead of ERA5.

These findings are discussed in the revised Section 3.3 (lines 393-403), where we discuss the precipitation forecast performance comparisons using IMERG as the reference dataset.

• Section 3.4: I remarked in my initial review that I appreciate single-observation experiments as they are informative about the sanity of the system and the information contributed by observations to analyses.

The choice of variable - in this case humidity - is ok, but it only shows the change of the humidity analysis in response to moisture-sensitive radiance perturbations. In data assimilation, one would then look at the response of the wind fields because moisture changes require advection. The change in winds would therefore indicate that the moisture changes happen for the right reason. If this is left out, the test carries little information on the physical consistency of the system.

There are still two 'increments' left in lines 416 and 425.

I still think that Section 5.2 on Jacobians in the Supplementary Information is too long.

This is known information and may only require a single figure, a paragraph and references.

Response:

The choice of variable - in this case humidity - is ok, but it only shows the change of the humidity analysis in response to moisture-sensitive radiance perturbations. …

We thank the reviewer for the valuable suggestion to include the changes in wind vector analysis fields in the single observation tests. And we have added a new Supplementary Figure 9 in Section 5.1 of the Supplementary Information. This figure illustrates that

the perturbation enhances the advection of drier air into a more humid region near the perturbed location, consistent with the results shown in Figure 4 of the main text.

Corresponding descriptions have been added to Section 3.4 of the main text (lines 444-449) to explain the newly included Supplementary Figure 9.

There are still two 'increments' left in lines 416 and 425.

We have fixed the inaccurate use of the term "increments" (lines 432 and 441).

I still think that Section 5.2 on Jacobians in the Supplementary Information is too long.

As suggested by the reviewer, we have only kept Jacobians for ATMS aboard NOAA-20.

• A general conclusion from the experiments with and without backgrounds is that the assimilated observational data is clearly not enough to constrain the analysis sufficiently well. This is mentioned in the Discussion but should also finds its way into the Abstract.

Response:

As suggested by the reviewer, we have revised the Abstract to include a statement highlighting the value of incorporating background forecasts in data assimilation.

3rd Review of "FuXi Weather: A data-to-forecast machine learning system for global weather"

Reply to Reviewer #1

We appreciate the reviewer's meticulous review and valuable feedback. We have made careful amendments to our paper, taking into account all the suggestions provided. Our responses to the comments, which are quoted in italic font, are detailed below.

Reviewer #1 (Remarks to the Author):

The authors did a lot of work clarifying parts of the manuscript based on reviewer suggestions, which I appreciate. I have a few remaining comments. I appreciate the authors responses and modifications to the paper, and I think the paper is nearly ready for publication. I have only a few minor comments remaining.

Comments

1. *Lines 161-163: consider including the Slivinski et al (2025) citation here.*

Response:

Following the reviewer's suggestion, we have added Slivinski et al (2025) citation in line 156 of the main text.

2. *Lines 395-396: consider replacing "indicating higher accuracy" with "relative to IMERG".*

Response:

As suggested by the reviewer, we have revised the original lines 395-396 (now renumbered as line 341) by replacing "indicating higher accuracy" with "relative to IMERG".

3. *Supplement: I think the added sentences on lines 94-101 regarding the 8h window would be more natural in the DA section instead (e.g. line 348).*

Response:

We appreciate the reviewer's suggestion and we have moved the added sentences originally in lines 94-101 to lines 343-351 in Section 2.2 of the Supplementary Information.

4. *Supplement lines 342-345: This is inconsistent with the statement that the assimilation window is 8h. These lines describe a window that is 7h long.*

Response:

Thank you for your insightful comment regarding the description of the assimilation window. We acknowledge the inconsistency and have revised the text to provide a clearer explanation. Specifically, satellite observations at a given time (e.g., 18 UTC) correspond to measurements collected within a one-hour window centered on that time, spanning from 17:30 to 18:30 UTC.

To enhance clarity, we have revised the relevant sentence in the Supplementary Information (lines 336-339) to: "… each within an 8-hour assimilation window that starts 3 hours before and ends 4 hours after the forecast initialization time, with each observation representing conditions over a one-hour interval centered on the observation time."

5. *Supplement, added Section 9: I appreciate the authors added this section. However, my previous comment was more a request for the technical details of this "direct from obs" experiment. Was a different model trained for the experiments without background fields? Does the architecture of that model differ from the one with background fields? Or was the same model used, but with fewer inputs than it was trained to take? Maybe I missed it, but it seems like these details would be useful for interpretation of the results that background fields are helpful for forecast accuracy.*

Response:

Thank you for your insightful comment regarding the need for additional technical details. In the "direct from obs" experiment, we used a separately trained model that does not incorporate background fields. This model shares the same architecture as the one used in experiments with background fields. Following the reviewer's suggestion, we have added the following clarification to lines 679-680 of the Supplementary Information:

"The two variants of FuXi-DA are trained independently but share the same model architecture described in Section 2.2."

Review of 'FuXi Weather: A data-to-forecast machine learning system for global weather' by X. Sun et al., submitted to Nature Communications, NCOMMS-24-80174-T

Summary:

The authors present an evolution of their machine-learning (ML) based forecasting system emulator Fuxi that is assimilating selected satellite based observations to create initial conditions for forecasts instead of using external analyses for initialisation. The data assimilation replacement method, called Fuxi-DA, is combined with an enhanced Fuxi forecast emulator to create the end-to-end Fuxi-Weather system that can be cycled like traditional, numerical methods based analysis-forecast systems. The output is assessed using standard forecast error metrics, and a specific focus is applied to a region in central Africa with sparse observational data coverage to test the robustness of the method.

The paper is complete regarding the explanation of methods, data, experiments and evaluation. The Fuxi developers build on an extensive record of ML applied to weather forecasting, and they exhibit deep knowledge of the methodological frameworks applied in the numerical weather prediction (NWP) community. They have also published their results before and made code available for others to try - which is highly commendable. The paper is very well written.

In the following, I have raised several points that will need addressing before publication. My greatest concern is the lack of evidence of statistical significance of forecast performance when compared against other data. More information is needed on the use of observational data in the data-assimilation process. I also wonder about the future need of data assimilation as an intermediate step between observational data records and forecast generation. Lastly, while I appreciate a special focus on central Africa for the reasons explained in the paper, the evaluation of performance for this case is not overly convincing.

General comments:

- *Need for data assimilation:* Given the recent evolution of ML-methods applied in weather prediction, I wonder whether there will still be a need for an explicit, intermediate data assimilation step in the future. The present need for this step originates from numerical models requiring starting conditions to evolve the differential equations in time. ML-methods do not necessarily need this unless they cycle through time steps like traditional systems. Recent studies seem to suggest that it may be advantageous to produce predictions directly from past and present observations, and therefore skip the data-assimilation step all together (aka direct observation prediction, e.g. `https://arxiv.org/abs/2407.15586`).
  While this point does not negate the importance of assessing ML-methods for data assimilation per-se, this paper could benefit from a discussion of this topic given that the target journal is Nature Comm.

- *Statistical significance:* It is standard practice in NWP evaluation to indicate the statistical significance of skill (differences). If the individual variance of two skill measures is too large, it will not allow to make statements about the significance of their difference: "Uncertainty in the measures can be indicated as confidence intervals or error bars calculated using approximations to statistical distributions, or through computer-intensive techniques such as re-sampling" (`https://empslocal.ex.ac.uk/people/staff/dbs202/publications/booksreports/mailier.pdf`).
  This applies to all figures 2, 3 and extended figures 3, 4, 5, 8, also maps. The interpretation of these figures will have to be adjusted as well as several score differences may not lead

to conclusions on better or worse performance. From my experience, some of the stated score differences will become meaningless once information on statistical significance is added.

- *Observation pre-processing:* The satellite radiance observations are 'assimilated' in a rather unqualified way. All channels of the microwave sounding instruments are used even though the window channels of both temperature and humidity sounders are very sensitive to surface emission that is driven by emissivity (dependent on surface type, moisture, vegetation, roughness) and surface temperature. These channels are also sensitive to clouds and precipitation, which greatly dilute the information content on the five target state variables. The authors mention cross-track scanning instrument (scan angle) biases. Are these corrected or is it assumed that - through knowledge of the scan-angle - the ML-method adjusts its weights? Can this be shown?
  The GNSS radio occultation data is used as profiles of refractive index, and this product exhibits larger errors in the lower 2-3 km of the troposphere and in the upper stratosphere/mesosphere. Pre-processing is mentioned, but the paper does not provide information on how systematic errors associated with this data are being addressed.
  Overall, the study uses observational data in a brute-force fashion that could be greatly improved on given the available knowledge from traditional data assimilation.

- *African case study:* I understand the motivation of the regional focus when applied to an area that is important for society, sparsely observed and where traditional system do not perform well. However, the study does not spend enough effort to fully assess the African case. Fig. 3 shows performance, again without statistical significance information, a two-cycle/day skill-score pattern (?), forecasts are verified against analyses that are notoriously poor in this area as opposed to (the few existing) observations, and for parameters that matter less for society here than, for example, soil moisture and rainfall. This is not a very convincing.

Specific comments:

- L56-61/L152-154, and throughout: See General Comments on performance outcomes given statistical significance above, and from selected comments below.

- L89: Please also mention efforts for ensemble forecasts towards uncertainty quantification for both traditional (which also feed off ensemble analyses) and ML methods.

- L207-212: Fuxi forecasts are evaluated with ERA-5, but ERA-5 is based on ECMWF's operational system from 2016 (nearly 10 years ago!), it has fairly coarse resolution (nominally 31 km, but analysis increments are produced at 124 km!).
  HRES seems to be evaluated with HRES at t0. This will favour HRES, and also affect the regional assessment over Africa with strong orographic variability that ERA-5 will not resolve (, and for which MSL is not the best parameter to choose by the way). RMSE-evaluations may actually suffer from higher resolution as increased variability penalises RMSE.
  Further, the NWP community does indeed evaluate forecasts with both own and reference analyses. Ideally, you show both and include HRES vs ERA-5.

- Fig. 2 and others: I assume the dip in forecast skill for Fuxi at day 4 is explained by the switch from Fuxi-Short to Fuxi-Medium. This information could be added in the caption.

- Section 3.1: Would it not be informative to compare global analyses fields to begin with, like Fuxi-DA vs ERA5 vs HRES at t0 for MSL, t2m and u/v300? This will exhibit, for

example, the analysis activity = standard deviation / climatological mean, which is important for RMSE evaluations, but also mean differences, for example, systematically colder/warmer or windier/calmer analyses.

- Line 286-288: Global model performance is actually evaluated for regions, for example, Europe, North America, East Asia, Australia etc. While this is not tied to socioeconomic implications, it is still a regional evaluation.
- L308-310: So where does the enhanced skill of Fuxi come from if ERA-5 is troubled over less well observed regions, surface in-situ observations are missing, and satellite data is affected by all kinds of signals from lower atmosphere, clouds, precipitation and surface characteristics (see above)?
- L312: 'performed or 'outperformed'?
- Section 3.4: I like the use of single-observation experiments. For clarity: are you showing increments or increment differences between the analyses from using a perturbed minus an unperturbed ATMS radiance? It would be great to see whether a 3D-Var would show the same increments or their differences.
- Supplement L190-193: If not treated properly (see General Comments item no. 3), the error characteristics from different instruments on different satellites become merged when averaging to the Fuxi grid!
- Supplement L211.1: I am not sure I fully understand how the PointPillars method is applied to GNSS RO data. Can the 3d-aspect of this be explained with a figure?
- Supplement Section 5: This section seems too long, and 3D-Var/Jacobians are standard knowledge. I suggest to reduce this section and focus on the single-observation experiment set-up, a summary of the results and in how far they deviate from, say, 3D-Var.

2nd Review of 'FuXi Weather: A data-to-forecast machine learning system for global weather' by X. Sun et al., submitted to Nature Communications, NCOMMS-24-80174-T

Summary:

The authors revised their paper in response to three reviews. The clarity of the paper has greatly improved but I still feel that it needs several corrections and changes before publication.

Comments:

- Abstract: ECMWF would claim that they are running an all-grid, all-sky/surface, all-channel forecasting system, which Fuxi is actually benefiting from because ERA-5 supplies the training data. ECMWF's exclude selected observations because of too large uncertainties due to cloud/precipitation/surface contamination. However, the fact that Fuxi includes all radiometer channels in their training (in a rather unqualified manner) does not mean that all channels contribute useful information in all situations. Fuxi would have to mimic ECMWF's observation screening to understand whether the more aggressive approach is beneficial or not.

- Section 3.2 and Fig. 3: Fuxi seems to outperform ECMWF mostly for days 6 and beyond for the important variables, namely geopotential height, temperature and wind. The more general statements on Fuxi outperforming ECMWF in the Abstract should therefore be more specific.
  I really appreciate the addition on analysis and forecast activity that was also requested by Reviewer no. 2. However, when looking at Fuxi's forecast activity, even after its enhancement, it is still substantially lower than that of ECMWF. This happens to coincide with the forecast range where Fuxi appears to be better. A sound conclusion could be that Fuxi only beats ECMWF forecasts at forecast ranges where it is quieter and will therefore produce better RMSEs.
  The claims on performance in both Abstract and Discussion need to be very clear on this.

- Section 3.3: I still think that verification against analyses in these areas is not very useful, and that a verification against observations would carry a lot more weight!

- Section 3.4: I remarked in my initial review that I appreciate single-observation experiments as they are informative about the sanity of the system and the information contributed by observations to analyses.
  The choice of variable - in this case humidity - is ok, but it only shows the change of the humidity analysis in response to moisture-sensitive radiance perturbations. In data assimilation, one would then look at the response of the wind fields because moisture changes require advection. The change in winds would therefore indicate that the moisture changes happen for the right reason. If this is left out, the test carries little information on the physical consistency of the system.
  There are still two 'increments' left in lines 416 and 425.
  I still think that Section 5.2 on Jacobians in the Supplementary Information is too long. This is known information and may only require a single figure, a paragraph and references.

- A general conclusion from the experiments with and without backgrounds is that the assimilated observational data is clearly not enough to constrain the analysis sufficiently well. This is mentioned in the Discussion but should also finds its way into the Abstract.