Reviewer Response

We sincerely thank the reviewers for their thorough evaluation of our manuscript. We appreciate the significant time and effort dedicated to providing such detailed and constructive comments, especially considering how we can best communicate with our target audience. We believe this has substantially improved the quality and clarity of our work.

Reviewer comments are written below in black. Our response in blue and changes to the manuscript are included in "quotation marks".

Review 1:

Summary: This paper highlights key biases across domains in Earth System Modelling. It is presented in a timely manner as we anticipate the next generation of CMIP model simulations. The paper does not put forth new scientific ideas but is a review of the literature on model biases and uses diagnostics from the CMIP AR7 Fast Track Rapid Evaluation Framework as exemplars. The paper thus serves to identify science gaps for future model development.

Comments:

Major:

1. A reader of this paper, a few years down the line may wonder why there is not a single figure describing any of the biases discussed in the paper. For that reason, it might be a good idea to not only motivate why this paper was written but why it was written this way - the last lines of the introduction give some idea idea but the title doesn't quite do this. The paper as such is a review/synthesis and does not actually diagnose biases across CMIP generations, so it would be good for the title to reflect that.

We have changed the title to "Synthesising" biases instead of diagnosing biases and made the review aspect of our paper more clear in the introduction. We aim to highlight that including diagnostics from across the Earth system is a unique aspect of our review. Moreover, we have now included several key figures from the literature.

"Here we present the community survey results used to select REF diagnostics, and synthesize literature from CMIP6 and earlier CMIP experiments to explore the origins and consequences of systematic model biases across multiple components of the Earth system, including several of the longstanding biases identified by Stouffer et al. (2017). We bring together existing literature to reveal how these biases interact across land, ocean and atmosphere. Where possible, we review how biases have changed between CMIP6 and earlier CMIP phases, and potential solutions to reduce these biases in CMIP7 and other future experiments."

2. Line 894: Not sure if this sentence was intended to be written as it is now as it doesn't quite read correctly. However, the statement about high-ECS models being excluded is incorrect. Please see Chapter 7 (Section 7.5) of IPCC AR6 WG1. In particular note that it says : "Solely based on its ECS or TCR values an individual ESM cannot be ruled out as implausible, though some models with high (greater than 5°C) and low (less than 2°C) ECS are less consistent with past climate change (high confidence).". The reasons for not using ECS and TCR directly from models in AR6 is also clearly outlined. It would therefore be incorrect to make the claim that high-ECS models were excluded (as the sentence currently implies) since the IPCC AR6 does not back that up. See Swaminathan et al, 2024 (https://doi.org/10.1029/2024EF004901) and Appendix C

in Dunne et al, 2025 (https://doi.org/10.5194/gmd-18-6671-2025) for a discussion on why ECS alone should not be used as a basis for model sub selection for assessments. The discussion in the TCR section on this is noted but perhaps worth rewriting the conclusion in the ECS section as well.

Thanks for pointing this out and the detailed comment including the linked articles. The previous statement was inaccurate we have removed that sentence and added a paragraph about what AR6 actually did, which is to move away from using climate model derived ECS and TCR entirely and use multiple lines of independent evidence to form their assessed ECS range and future warming projection instead. In addition, we have also added a couple sentences on the different perspectives on how to handle models with high ECS.:

"The inconsistency between modelled and historical warming is one of the main reasons why the IPCC AR6 report moved away from using ECS and TCR values derived directly from climate models and instead used multiple lines of independent evidence based on instrumental records, paleoclimate evidence, emergent constraints and improved process understanding to form their projected warming (?). Another reason is because climate models are already used as a tool to support these different lines of evidence, such as in estimating global temperature in paleoclimates and improving process understanding of different climate feedbacks. Because the AR6 assessed ECS is lower than the CMIP multi-model mean ECS, the global surface warming in AR6 is lower than the raw model mean. This leads to disagreements about how high ECS models should be treated. Some authors suggest that 'hot models' should only be included in considering the impacts at specific levels of global warming (largely independent of time or scenario) to avoid the issue of these models projecting a much warmer future (?). Whereas some authors argue that high ECS models should not be discarded as their high ECS is sometimes based on realistic improvements in cloud processes (Zelinka et al., 2020; Bodas-Salcedo et al., 2019) and therefore can provide physically consistent assessment for future impacts and risk assessments (Sutton, 2019)."

Minor:

1. In the Abstract, the word "hypothetical" in the first sentence suggests that the scenarios could be imagined and not really true. Suggest using a phrase with scenario in it which is more the scientific terminology and appropriate.

We have changed this to 'idealized forcing scenarios'.

2. If possible, it would be good to not have acronyms or at least acronyms that are not expanded in the Abstract. This would be helpful for readers outside the CMIP community.

Thanks for pointing this out, this has been changed in the abstract.

3. Line 21: The phrase "dozens of unique" may not be factually correct. Some models share components and if one counts just the unique models, they may not be dozens. Maybe providing some approximate numbers instead would add weight to the claim that there are many models including many independent ones?

This sentence has been changed to give more precise background information to the reader:

"The most recent iteration, CMIP6, comprises dozens of nominally distinct coupled climate models, including both General Circulation Models (GCMs) and Earth System Models (ESMs), which performed standardized historical, future, and idealised simulations to advance understanding of the Earth system and its evolution. These models can be traced to around 14 broad model code families (Kuma et al., 2023), each of which may contain several models that have since undergone substantial independent development."

4. Line 24: Just a caution that in CMIP6, scenarios do not just include forcings but socio-economic pathways too. This should be made clear.

Thanks. We have added this.

5. Line 32: Should that be "interactive" carbon cycle and not "active"? I think that might be the term used in the literature. (as opposed to prescribed).

Thanks, you are correct.

6. Line 35: Suggested change to make it grammatically better - One aim of the CMIP6..was to answer the question - "What are.."

Thanks. We have changed this.

7. Line 38: Note to change the REF paper reference as the paper is now published.

We have updated the reference throughout the paper.

8. Line 40: Should this be a reference to the Data Request paper as this is more a reference to where the 5 domains originated from? This might be https://egusphere.copernicus.org/preprints/2026/egusphere-2026-1641/

We refer to "The first set of fast track REF diagnostics " here rather than the origin of the 5 domains. Hoffman et al. (2026) is the correct reference.

9. In Table 1 : Note that the final list of diagnostics in the REF paper has been revised for various reasons. Perhaps putting an asterisk where it says "final" and explaining this was a post-survey summary might clarify the discrepancy between this list and the REF paper list.

Thanks for noticing this. We have changed the legend to: *"The diagnostics and themes chosen for priority evaluation in the REF following the survey. Note that the current REF diagnostics have been further adapted (Hoffman et al., 2026).* "

10. In Table1, Also, some of the numbering seems off and Snow cover doesn't have a number. If there is a reason for this, perhaps an explanation can be provided so as to not confuse the reader.

Some diagnostics were not included as they were considered too broad. Other diagnostics were grouped so several diagnostics will share the same section number. We indicate that in table now.

11. Line 81: Suggest changing "these both" -> "Both these".

This has been changed.

12. In Figure, 3 could the authors clarify how the boxes for Key Diagnostics an Other Diagnostic Processes differ? It is not clear how that legend item is used.

"Key diagnostics included in this review (rounded corners) and other diagnostics or processes that drive biases in the Ocean (flat corners) are grouped to show interaction across diagnostics and themes: Ocean and sea ice (blue shading), land and land ice (yellow shading), atmosphere (green shading), Earth system (red shading) and mixed themes (grey shading)."

13. Line 144: Perhaps not quite correct to call these "inaccurate", they can be "inadequate" or some other term that the authors can think of. Inaccuracy can be read as models doing something wrong but that's not quite the case.

Indeed, we now use "deficient cloud cover representation"

14. Line 136: It is not clear how the authors are linking biases in SST to reasons through Figure 3. It would be helpful to call attentions to specific boxes and highlight some of the reasons. Suggest that this should be replicated for other figures as well to help drive home the message from this manuscript.

This is a really good point and we have endeavoured to reference the figures better. We now begin each section with a 2 sentence overview of the figure and reference specific boxes throughout the manuscript.

15. Line 216: Suggest removing the word "inaccurate" as "error compensation" conveys the message here.

We agree. This has been changed.

16. Line 294: Perhaps the authors could qualify that most CMIP models do not have interactive Greenland Ice sheets. UKESM 1.2 (https://egusphere.copernicus.org/preprints/2025/egusphere-2025-4476/egusphere-2025-4476.pdf) and I think IPSL have interactive Greenlan ice sheets. There doesn't appear to be a documentation or reference paper for IPSL yet though.

We rephrased to address this as CMIP models did not include interactive ice sheets for their historical runs. While several have the capacity for CMIP7, its unclear how many will use it:

"Standard CMIP configurations[1] do not include an interactive Greenland Ice Sheet (GrIS) component, limiting their ability to represent variability and trends in freshwater fluxes associated with ice-sheet mass loss and their potential influence on the AMOC. [1]Only very few CMIP6 models include interactive ice sheets (e.g., UKESM1-ice-LL (Smith et al., 2021) and CESM2.1-CISM2.1 (Muntjewerf et al., 2021))"

17. Line 373: I am not sure how we would characterize or quantify all ENSO teleconnection biases since we may not even be fully aware of all possible teleconnections? Would the authors consider including this aspect of teleconnections as well?

Thank you for raising this point. In our synthesis here for ENSO teleconnections, we primarily focus on ENSO's impact on regional temperature and rainfall based on the findings from Planton et al. (2021) and Lee et al. (2021), and now we have clarified this. ENSO rainfall and temperature teleconnections are quite well understood in existing literature; however, there are limitations around evaluating the strength of these teleconnections, given their non-stationary nature due to internal variability, and we have now explicitly highlighted this. As you have correctly pointed out, there are many ENSO teleconnection biases that are difficult to quantify or characterise partly because our we lack in our physical understanding of them. We have now highlighted some examples of these such as inter-basin interactions between ENSO and NAO; ENSO's impact on other modes of variability; and ENSO's impact on circulation processes such as AMOC. We are not including these in our synthesis mainly due to lack of literature around these teleconnection biases (particularly for early generation coupled models) and also due to incompatibility with how ENSO teleconnection diagnostics are defined in the REF and other past studies on ENSO metrics. You can find the changes in section 3.1.4 of the revised manuscript:

"ENSO teleconnections are also non-stationary in nature and can vary significantly on decadal to multi-decadal timescales in multiple regions around the world (Gulev et al., 2021; Miao et al., 2025). As a result, evaluating the representation of regional teleconnection strength in models over a short historical record is extremely challenging. In addition, there are other aspects, such as ENSO's impact on other tropical modes of variability like the North Atlantic Oscillation (NAO) (Kim et al., 2026; Scaife et al., 2024) and key circulation patterns like AMOC (Wunderling et al., 2024), which are less well understood and still an active area of climate research. For this section, we primarily classify ENSO teleconnections as ENSO's influence on regional precipitation and

temperature patterns alone (Planton et al., 2021)."

"ENSO teleconnections are also non-stationary in nature and can vary significantly on decadal to multi-decadal timescales in multiple regions around the world (Gulev et al., 2021; Miao et al., 2025). As a result, evaluating the representation of regional teleconnection strength in models over a short historical record is extremely challenging. In addition, there are other aspects, such as ENSO's impact on other modes of variability like the North Atlantic Oscillation (NAO) (Kim et al., 2026; Scaife et al., 2024) and key circulation patterns like AMOC (Wunderling et al., 2024), which are less well understood and still an active area of climate research. For this section, we primarily classify ENSO teleconnections as ENSO's influence on regional precipitation and temperature patterns alone (Planton et al., 2021)."

18. Line 717: Suggest changing "culprit" to phenomenon or feature as that would be more appropriate.

Agreed and we have changed it to "feature".

19. Section 4.2 : Suggest including a recent publication from the CMIP7 Model Benchmarking Task Team that discusses observational uncertainties in model evaluation:
`https://doi.org/10.1175/BAMS-D-25-0079.1`

Thank you for pointing out this publication. It is highly relevant to our discussion. In addition to changes in Section 4.2, we have drawn from it in two further places in the Conclusions.
In Section 4.2, we have strengthened our discussion of why reference datasets are themselves uncertain and we now direct readers for a more comprehensive overview and for recommended best practice:

"Moreover, many 'observed' global-scale products in fact undergo some processing, and in some cases the observed and simulated variables do not represent the same physical quantity despite sharing a name, for example, leaf area index is defined differently across communities (Beadling et al., 2026)."

"*Beadling et al. (2026) discuss these issues in detail and recommend best practices for handling observational uncertainty in CMIP-era model evaluation, including the use of multiple independent reference datasets and clearer documentation of uncertainty by data providers.*"

We also found two points in this publication sufficiently relevant to our perspective on CMIP7 so we have incorporated them into the Conclusions. First, our discussion of machine learning previously considered only hybrid parameterizations but now we note that machine learning is increasingly applied on the observational side as well, and that comparable caveats exist for that:

"*Machine learning is also increasingly used in the construction of the observational products against which models are evaluated, including the upscaling of point-source measurements to global grids, spatio-temporal interpolation and gap filling (Beadling et al., 2026). While these methods help address data scarcity, they can introduce uncertainties that are difficult to quantify and may not faithfully reproduce interannual variability, implying that the same concerns regarding generalisability and physical consistency apply to reference datasets as to parameterizations.*"

Second, we have added that the evaluation benefits of higher-resolution and higher-frequency simulations depend on a parallel expansion of observational capability through new campaigns, and continuation of sustained in-situ measurements:

"*Realising the benefits of moving to high-resolution model simulations will require a parallel*

expansion of observational capability. High-resolution and higher-frequency simulations can only be assessed against observations of comparable resolution and frequency, which remain scarce over the long records needed to characterise extremes, while sustained in-situ measurements of the subsurface ocean, which cannot be obtained from satellites, are vulnerable to funding lapses and instrument retirements (Beadling et al., 2026). Continued investment in new and sustained observational campaigns, particularly in data-sparse regions, is therefore a prerequisite for evaluating the next generation of models."

Review 2:
Decision: major revisions needed.

The submitted manuscript presents an overview of common climate model biases that have been documented and persistent across multiple CMIP generations. Specifically, the authors focus on variables (diagnostics) that have been selected for assessment by the Rapid Evaluation Framework (REF). The paper aims to provide a literature review of the performance across CMIP generations for the REF diagnostics and is aimed at endusers who may be new to CMIP or are looking to connect realms (e.g., understand consequences or relevant of biases in the ocean and their connection to ice, atmosphere, land, etc.). I recognize that what the authors are trying to achieve is a large and valuable task, so I commend their efforts taking this on. Specifically, it is very important to clearly summarize model biases and the consequences of them to end users who might not have domain expertise but wish to use CMIP results for various applications. Therefore, I think the topic is important to the community. However, in its current state and given its goals and intended audience, I believe this manuscript needs substantial revision prior to being ready for publication. I have made many comments below in my review (I have reviewed the entire manuscript, but my comments are mostly focused on the ocean and ice portions, where my expertise lies).

We thank the reviewer for their detailed comments ad we have substantially revised the text. An overview of changes are described below but we have revised each section to address the major points raised and these can be seen in the track changes document.

Overall, given that the authors state one of their goals is to present a literature review, I felt that several topics were incomplete in their assessment of relevant literature and would greatly benefit from more discussion.

- The entire paper would also greatly benefit from attempting to quantify and discuss biases from CMIP3 to CMIP6. Much of the literature discussed, I felt was focused on CMIP6 and there was limited discussion of cross-generational performance / improvements. We have included many more references to CMIP3 and CMIP5 throughout, showing improvements across generations as well as persistent weaknesses. Changes have been made across most sections and diagnostics.

- There were also limited discussion / examples of "success stories" in bias reduction across generations. Can the authors identify any within these realms? Thanks for highlighting this. We have taken this on board and added several more examples of successes. Some examples:

- We have highlighted some regional SST bias improvements between CMIP5 and CMIP6, along with some regional SST biases that have got worse;

- We have summarised the improvements in ENSO and what to look forward to;

- We have highlighted improvements in synoptic weather representation;

- We have highlighted the role of nitrogen limitation in land surface representation;

- For the AMOC, we have added two examples of progress: reduced meridional heat-transport biases across CMIP3, CMIP5 and CMIP6 , and improved representation of interannual variability in CMIP6 compared with CMIP5.

- There are also issues throughout the text regarding accuracy and clarity of some statements, which I have noted in the line-comments below.

- I also strongly believe that the manuscript would greatly benefit from the inclusion of a set of figures for each diagnostic discussed to illustrate the biases across the CMIP ensemble. This should be feasible, especially since most of these diagnostics are relatively lightweight to compute and the authors could even leverage REF software / code if necessary. This would greatly increase the impact and readability of the paper, as right now it is very hard to follow the discussion without maps and visualizations to illustrate the issues at hand.

  We have added several more figures from literature and made ourselves to address this point. The figures we have produced are:

- ENSO biases in CMIP5 and CMIP6

- The performance gain in simulating synoptic weather types from CMIP5 to 6 in the subsection "Synoptic Weather Types", calculated with the the *pyLamb* package available from `https://doi.org/10.5281/zenodo.15346362`

- Soil Moisture bias.

- A schematic of AMOC biases.

The methodologies are described in Appendix B. Additional figures from the literature can be found in the text.

Line-by-line and minor comments can be found below:
Abstract:

"The Coupled Model Intercomparison Project (CMIP) **simulates** standardized experiments spanning historical, future, and hypothetical conditions, to better understand the Earth system's evolution." -> **I suggest using another word than "simulates" here as CMIP itself does not "simulate" anything. CMIP "provides the framework to standardize, organize, and archive results from . . . ".**

This has been changed to:
"The Coupled Model Intercomparison Project (CMIP) coordinates standardized experiments spanning historical, future, and idealized forcing scenarios, enabling systematic investigation of the Earth system's evolution."

"Systematic biases remain" -> **Systematic biases of climatically relevant processes . . . all models will have biases . . . the limitation is when they are in regions / for processes that are important to the climate system (e.g., by controlling rates of ocean heat uptake, carbon budgets, etc).**

Thanks for this comment. We have added it.

"AMOC´s large scale transport biases" -> **apostrophe is off in AMOC. Also, since the sentence already starts with "biases", you can remove "biases" here after "transport".**

This has been changed.

L12: "enhanced coupling" -> "improved coupling" ?

This has been changed..

L13: "the current state of understanding" -> drop the "s" after "understandings"
This has been changed.

L18: "almost 30 years", -> It is officially over 30 years as the framework began in 1995 (AMIP started in 1989). See https://climate.esa.int/en/news-events/cmip-the-evolution-of-climate-modelling/
This has been changed.

L22: "historical, future, and experimental conditions," -> I understand what is trying to be conveyed by "experimental conditions", but it is a bit confusing to a reader not familiar with CMIP .... Future scenarios are also "experimental conditions". Can you please reword and be more explicit here by what you mean?
Great point. We have changed to "idealised".

L23: "experience as a function of different climate forcings." -> I suggest "as a result of different possible future emission pathways" ... it really is emission pathways (which come with a lot of other societal assumptions, etc ... not isolated forcings, in the case of SSPs. Can you also please add a citation here?
We have been more explicit: *"...may experience as a function of different climate forcings and socio-economic pathways (O'Niell et al., 2016; van Vuuren et al., 2026)."*

L25: "Through all the phases of CMIP phases, there have been systematic model biases, where the behavior of **climate phenomena simulated in** historical experiments deviate from observations of the Earth system." -> It is not that the behavior of the historical experiments themselves differ .... It's that certain climate processes within the simulations differ. You can have a model that does a good job at the trajectory of global ave SST or SAT (which we do well quite well!), but still has very wrong sea ice representation, for example.
We have added this as you suggest.

L27: "higher extension" -> please add "vertical".
We have done this.

L33: "The focus of CMIP7 on emission-driven rather than concentration-driven future scenarios will **increase the effect** of how well the carbon cycle is represented in coupled climate models." -> **what does "increase the effect of" imply here? Running in emission driven vs. concentration driven doesn't necessarily improve *how the carbon cycle is represented***. Please reword for accuracy. More comprehensive ESMs are what improve carbon-cycle representation (e.g., more biogeochemical tracers in the ocean for example, or atmospheric chemistry (prescribed vs. interactive ozone) ...). If models have biases in their physical components, running in "emissions-driven" mode and allowing the feedbacks between simulated climate components can actually make things worse! The ocean and atmosphere carbon stocks can end up being out of sink with the observed rate of atmospheric CO_2 increase.
We mean to say there will be greater attention and greater consequences for the simulation: "*The focus of CMIP7 on emission-driven rather than concentration-driven future scenarios will will place greater emphasis on the representation of the carbon cycle in coupled climate models, while increasing its influence on simulated climate outcomes.*"

L38: "The REF platform will enable comparison between models and across idealized and historical experiments against observational evidences." -> **I believe it is only historical experiments, not idealized. Please check. Also, change "observational evidences" to "observational estimates" ...**

We would like to keep "idealized" in the text as hist-GHG is, for example, included and this is more of a sensitivity test than an accurate representation of the real historical conditions. Furthermore, the REF is likely to be updated and users can make their own modifications so may include idealised experiments. We have changes 'evidences' to 'estimates' as this is clearer.

L40: The Hoffman et al., reference can now be updated to the most recent.
Thanks for pointing that out. The reference was updated throughout the manuscript.

L44: "across all components of the" -> Across "multiple components" would be more accurate here as CMIP models do not include *all* components (e.g., ice sheets).
This has been changed.

L79: Please fix the apostrophe before S in Sea and elsewhere throughout this section.
This has been changed.

L122: "Temperature and salinity variations affect the density of seawater, controlling the ocean's stratification, lateral pressure gradients **which drive ocean transport**, and **the formation of water masses that fill the ocean interior**. This impacts ocean overturning and circulation at a range of scales." -> See my edits. I think it is very important to mention that stratification is essentially what controls water mass transformation / formation / subduction / transport **..... this is what is relevant for ocean heat and carbon uptake ....**
This has been changed.

L126: Add "processes" after "climate". Again ... I think it is important to state that air-sea fluxes set the interior ocean water mass structure and influence ocean heat uptake and ocean carbon uptake ...
This has been changed.

L130: Region AND model dependent! Some models have improved quite a bit!!!
Done.

L138: "particularly in regions **where observations are sparse** and model constraints are weak especially in the deep ocean **where observations are sparse". Please elaborate, a reader might not understand why biases in the deep ocean lead to SST biases?**

We have adjusted the text significantly based on this comment and your next comment, including a note on overturning as a route to propagate these interior errors to the surface (see next comment's response).

L140: "can introduce initial shocks and model drift". -> This is a very jargony / technical sentence for the target audience. I assume you are referring to "initialization shocks" that lead to "numerical / spurious drift". Please explain how everything you listed above can lead to model drift, I am not following. This sentence also needs a citation.
We have adjusted these sentences and moved a relevant citation to the later sentence on shocks and drift:

"*..., particularly in regions where observations are sparse and model constraints are weak (Huang et al., 2003; Hawkings & Sutton, 2009; Zheng et al., 2011; Zhang et al., 2023; Zhang et al., 2024). The ocean state used to initialize simulations is often based on sparse measurements, particularly in the poorly-observed deep ocean, and must also be interpolated onto the ocean model grid. Inconsistencies between this gridded initial state and the valid ocean states of the dynamical*

model can introduce initial shocks and persistent drift within the model solutions (Exarchou et al., 2018) which can manifest as biases in the ocean state including the SST as water circulates laterally and via overturning."

L144: "The persisting Southern Ocean warm bias in several generations of CMIP models is mainly attributed to inaccurate representation of AMOC (Wang et al., 2014), and clouds (Hyder et al., 2018)." **This is much too strong of a statement and relies on the results from only the two cited studies. The cause of the warm SST bias in the Southern Ocean is a topic of on-going research and other studies suggest it may also be due to errors in the surface freshwater budget potentially linked to the lack of glacial meltwater inclusion. Please provide more references for this and nuance.**

Figure 6 in Beadling et al., 2020 shows Southern Ocean upper-ocean T and S biases. A key take away: *"The upper ocean remains biased too fresh and too warm relative to observed. There has not been a clear improvement in ensemble performance relative to CMIP3 and CMIP5. There are still several models with exceptionally warm upper SOs with errors that translate into very poor representations of Antarctic SIE." https://journals.ametsoc.org/view/journals/clim/33/15/jcliD190970.xml*

We agree that the original attribution was too strong. We explicitly relate this freshwater evidence to historical SST trends rather than treating it as an explanation of the climatological warm bias alone. We have changed the text to include this nuance and additional references:
"The persisting Southern Ocean warm bias in several generations of CMIP models is an area of ongoing research, and has been attributed to inaccurate representation of AMOC strength (Wang et al., 2014), vertical mixing driven by internal waves (Song et al., 2025), large-scale ocean transport (Rackow et al., 2022) and cloud-related shortwave radiation (Hyder et al., 2018), or to warm biases in deep ocean water (Luo et al., 2023). For historical SST trends, underestimated increases in precipitation minus evaporation and missing freshwater inputs from Antarctic ice-sheet and ice-shelf mass loss can contribute to excessive simulated Southern Ocean warming (Kaufman et al., 2025) (Fig. **??***, Water Sources and Sinks). These Southern Ocean warm biases impact the ability of models to simulate Antarctic sea ice extent (e.g., Beadling et al., 2020)."*

L151: "distribution, circulation and". Are you referring to atmosphere or ocean circulation here?
We were referring to atmospheric circulation. This has been clarified in the text.

L159 – 160: "Ocean density and related thermodynamics are especially sensitive to salinity variations in polar regions where water is cold, compared to more temperate regions where density fluctuations are largely driven by temperature changes." -> **This is due to the non-linearity in the equation of state of seawater. Please add a citation. A reader from the target audience may not understand why cold temperatures mean a region is more sensitive to salinity (it is because the thermal expansion coefficient depends on temperature).**
We have added the following sentence:
"*This is due to the non-linear equation of state of seawater (relation between water density and its temperature, salinity and pressure), where the thermal expansion coefficient is temperature-dependant and becomes small in cold water (Stewart and Haine, 2016).*"

L166-167: ", and the Southern Ocean freshwater fluxes are underestimated in CMIP6 (Kaufman et al., 2025).". -> **Indeed, but more explanation is needed** – lack of glacial meltwater fluxes!! Also, a reader might be very confused by this discussion, the fact that you state the ocean is too fresh (also see Figure 6, in Beadling et al., 2020 linked above) but also that the freshwater fluxes are underestimated. Please draw the connection for the reader.

We have adjusted this section to account for this comment and the following comment. It now reads:
"In polar regions, the Arctic basin fresh bias was reduced between CMIP5 and CMIP6 but still persists (Wang et al., 2022), and the near-surface waters in the Southern Ocean are too fresh across CMIP3, CMIP5 and CMIP6 (Beadling et al., 2020). The Southern Ocean also suffers from an underestimation bias in freshwater flux (precipitation and glacial run-off due to Antarctic ice-sheet mass loss), which contributes to the Southern Ocean warm bias discussed above (Kaufman et al., 2025). The warm and fresh biases of the Southern Ocean are also related to biases in simulated Antarctic Sea Ice Extent, a critical property for air-sea exchange which is discussed later, although it is not clear which is the causal bias and which is the resultant bias (Beadling et al., 2020)."

L167: "run-off errors," -> what runoff errors are being referred to here? Too much runoff from Antarctica / Greenland? Please provide a citation that has linked the surface fresh biases to runoff and precip errors.
We have attempted to address this comment alongside the modifications to the prior comment (see above response and adjusted text).

L168: "The Indian Ocean is another region where SSS biases are important, with salinity stratification forming barrier layers and thin freshwater lenses near the ocean surface, mediating air-sea fluxes, and upper-ocean dynamics and thermodynamics." -> **the use of "barrier layers" is a bit awkward here, I suggest rewording. This statement also is not unique to the Indian Ocean.**
Thank you for the suggestion. We have reworded as:
"Sea surface salinity biases are also important in regions (often Equatorial) where salinity stratification can inhibit mixing between the upper ocean and the thermocline, mediating air-sea fluxes, and upper-ocean dynamics and thermodynamics."

L171: "with thick observed barrier layers, driven". **Please avoid using the term "barrier layers"**
Changed to: "*with thick salinity-stratified sub-surface layers*"

L174: "summarize" -> I am not sure that summarize is the right word here. Integrate?
Thank you for your suggestion. Changed to "*integrate*"

L176: "(warmed faster than observations)". -> Since you are talking about OHC, which is not just temperature, **but volume**, this should state "(has taken up heat faster than observed)". Since OHC is dependent on both temperature change and volume, where the heat goes matters for the full layer integral. Same comment on "cooling" used later in the sentence.

Changed text to:
"*Globally over the 1970-2020 period, the CMIP6 OHC exhibits a warming bias (i.e., has taken up heat faster than observations) near the surface (0 - 200 m) and below 2000 m depth, and a cooling bias (has taken up heat slower than observations) between approximately 400 - 2000 m depth.*"

L179 – 180: "This bias is attributable to inaccurate SST and biased mixing/surface fluxes into the warmest regions of the ocean" -> **What is "this bias" that you are referring to? The study by Sohail et al., referenced here indicates that the COLD bias is traced to SST errors in the warmest regions of the ocean. Please be precise.**

We adjusted the text following your suggestions. It now says:

"Globally, CMIP6 models exhibit contrasting OHC biases across depth ranges. Over the 1970-2020 period, the CMIP6 OHC exhibits warming bias (i.e., has taken up heat faster than observations) near the surface (0 - 200 m) and below 2000 m depth, and a cooling bias (has taken up heat slower than observations) between approximately 400 - 2000 m depth (Sohail et al., 2021). These opposing biases can partly cancel when OHC is integrated over broader depth ranges; integrated over the ocean's volume, there is an overall cooling bias across most CMIP6 models, which is attributable to inaccurate SST and biased mixing/surface fluxes into the warmest regions of the ocean (Sohail et al., 2021). The bias could be reduced through improved mixing parameterizations (Sohail et al., 2021). Meanwhile, averaging over just 0 - 700 m depth largely cancels out the opposing 0-200 m and 200-700 m depth OHC biases (Takano et al., 2023). "

L176 – 178: "near the surface (0 - 200 m) and below 2000 m depth, and a cooling bias between approximately 400 - 2000 m depth (Sohail et al., 2021). The opposing biases compensate and disappear when averaged over 0 - 700 m depth (Takano et al., 2023)." –> This sentence does not make logical sense as written. You mention that the opposing biases are in the upper 200 m (too much OHC) and over 400 to 2000 m (too little OHC), but then the compensation you mention is just within the upper 700 m. Please reword for accuracy.
Reworded (see text of prior comment's response).

L180 – 183: "In the Arctic, CMIP6 models exhibit a warm and fresh bias at depth through 1979-2014, with effectively no improvement in Arctic deep hydrography from CMIP5 to CMIP6. It was suggested that these Arctic biases could result from inaccurate mixing schemes, insufficient model resolution, and propagation of biases from adjacent ocean regions (Khosravi et al., 2022)." -> This sentence seems very out of place; this does not discuss OHC and instead discusses temperature and salinity biases.
This does seem out of place. We have moved this to follow the polar discussion within the sea surface salinity section. This seemed like the best-fit location for the sentence.

OHC Section overall: I felt that this was a very incomplete discussion of the literature regarding OHC and understanding where sources of these biases may come from. The Southern Ocean is the most important basin for the oceanic uptake of anthropogenic heat, but none of the relevant literature assessing OHC / OH uptake in this region was reviewed. It is very hard to follow the discussion without actual figures in the paper, the reader is required to go look at all the individual papers cited to understand where in the water column these biases are and whether or not they are significant.

We have rewritten the section to (i) be more clear and simplify the discussion around the different depths, (ii) include more literature and (iii) explicitly mention the SO.

Figure 3. An arrow connecting freshwater fluxes to "Sea Ice & Coastal Processes" should be drawn as freshwater flux biases impacts ocean stratification which influences sea ice formation processes and coastal heat bugets. Freshwater flux biases also impacts the formation of dense waters (e.g., dense shelf water formation along the Antarctic coast).
We agree. This arrow will be added in future revisions (although we did not have time to do so before this resubmission, as the figures must go through our design team).

L193: "It also defines the albedo (reflectivity) of the ocean surface . . . " -> I would reword for accuracy, **sea ice influences planetary albedo**, it is not part of the ocean surface, it sits on top of the ocean surface.

Thanks for pointing that out. We revised the sentence.

L193: "short-wave" should not be hyphenated here.

We changed that.

L194: "A loss of sea ice decreases **the** albedo, increases the heat uptake of the ocean" -> please be specific about what components of the heat budget you are referring to and that it increases "the planetary" or "the surface" albedo, not just "the albedo".

We changed the sentence and expanded a bit:
"A loss of sea ice decreases the surface albedo locally, which enables shortwave heating of the ocean that would otherwise be covered by sea ice, resulting in a positive ice–albedo feedback, which eventually leads to enhanced warming in polar regions (Forster et al., 2021). A correct representation in coupled climate models is essential for predicting future climate conditions, especially since both the poles show a strong transition to low sea ice states with less perennial ice and more seasonality (Spira et al., 2026)."

L204 – 217: There are more recent studies that can be discussed and referenced here which discuss the persistent of biases and potential sources / remaining uncertainties.

E.g., Roach and Polvani 2026 (`https://journals.ametsoc.org/view/journals/clim/39/7/JCLI-D-25-0477.1.xml`), Bushuk et al., 2026: `https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026GL124043`.

The paragraph in question discusses only mean-state biases, whereas both suggested papers deal solely in trend analysis. The suggested references are therefore more applicable to the following paragraph/s where they have been incorporated as requested.

L245 – 248: "When this trend component is included in comparisons with historical CMIP simulations, the long-term observed trend lies within the spread of modeled SIA trends, leading some to argue that models do in fact reasonably capture multidecadal Antarctic SIA trends overall (Holmes et al., 2024)." -> **Please read Roach and Polvani, 2026 and Bushuk et al., 2026 linked above which make a clear case against this statement (using a much larger ensemble of models). It is important that we do not propagate this narrative that models are "better at capturing" an observed trend .... a trend that is very non-linear!**

Thank you very much for pointing this out. While the ensemble does not seem to be larger, the paper by Roach and Polvani gives a counter perspective and definitely needs to be cited here. We revised this part of the manuscript accordingly. The section now reads:
"Some argue that when including the period of abrupt sea ice decline in trend comparisons with models (e.g. using ScenarioMIP data to encompass 1979-2023 rather than 1979-2014), models are more capable of representing average observed SIA trends (Holmes et al., 2024). However, others argue that the period of expansion and decline must be considered separately given that none of the models accurately represents them both (Roach and Polvani, 2026). Furthermore, the models that most closely reproduce the near-zero average observed trends have poor representation of the spatial distribution of those trends, as well as more generally of sea ice climatology and internal variability (Bushuk et al., 2026)."

L259: "AMOC remains one of the most biased large-scale ocean diagnostics in coupled models." **Relative to what? Boundary currents? The ACC? Gyres?**

The wording was indeed unprecise, we meant to refer to projections spread. We have reformulated to: *"Despite improvements from CMIP5 to CMIP6, substantial inter-model differences persist in the representation of the AMOC, contributing to a large spread in projected AMOC weakening and associated regional climate responses (Bellomo et al., 2021)."*

L286 – 288: "Heuzé (2021) also showed that CMIP6 models still struggle to represent Antarctic Bottom Water formation, with overly warm and fresh deep waters and insufficient

bottom-water export, highlighting persistent global deficiencies in deep-water mass formation and circulation (Fig. 3)." -> I suggest removing this sentence as AABW is driven by very different mechanisms and this sentences loses the focus on the AMOC.

Thank you for this comment. We agree that AABW formation is governed by different mechanisms from NADW formation. However, AABW contributes to the abyssal overturning cell in the Atlantic and is therefore relevant to the vertical structure of the overturning circulation. We have revised the sentence to clarify this connection and maintain the focus on AMOC-related structural biases.

L257: Please add "northern" before "latitudes".
Changed

—Section 4—-

Section 4 in general: I recognize that this section is supposed to serve as a bringing together of everything that was discussed prior, but there are several large claims being made without any citations.

Thank you for your comment. We have added relevant citations across Section 4.1

E.g., *"Parameterized mixing affects ocean stratification and deep-ocean heat uptake, contributing to widespread SST and salinity biases (Fig. 3), including the equatorial Pacific cold-tongue bias and Southern Ocean warm bias seen across CMIP6 models."*.

This sentence holds a lot of weight and requires a reader to understand why parameterized mixing might be linked to the Southern Ocean being too warm .... But in fact, even models that are "eddy permitting" in the mid-latitudes / latitudes of the ACC still show warm SSTs in the Southern Ocean ..... so that is not the full story.

As this is a summary we have removed these references that are too complicated to explain in one sentence:
"Parameterized mixing schemes influence ocean stratification and the vertical redistribution of heat and salt, affecting SST and salinity biases and deep-ocean heat uptake."

E.g., *"Systematic biases in sea ice area (SIA) in both the Arctic and Antarctic arise from the spatiotemporal distribution of ice cover, and the response of sea ice to anthropogenic warming. These discrepancies are associated with deficiencies in the representation of mass balance components, ice-albedo feedbacks, and the underlying physical processes that control sea ice cover. The omission or oversimplification of key processes in some models, such as ice-sheet-ocean interactions and freshwater fluxes, further reduces confidence in simulated sea ice dynamics and projections, and feed into biases in the ocean (Fig. 3)."*

E.g., *"Inadequate representation of cloud fraction, phase partitioning, and optical depth affects both incoming solar radiation and the trapping of outgoing longwave radiation, contributing to biases in land and ocean, such as warm biases in the Southern Ocean (Fig. 3)."*

E.g., *"Convection schemes often misrepresent entrainment rates, trigger thresholds, and convective organization, contributing to the double ITCZ problem, unrealistic monsoon variability, and shifted storm tracks."*

Please also break up this discussion (Sections 4.1 & 4.2) into distinct paragraphs for readability.

Perhaps the formatting is unclear but there are distinct paragraphs, hopefully this is clear in the final typesetting.

Section 4.2: You might want to consider referencing this recent paper that discusses what is being summarized: https://journals.ametsoc.org/view/journals/bams/107/4/BAMS-D-25-

0079.1.xml

Thank you for this suggestion, which was also raised by the other reviewer. We agree that this publication is highly relevant and we now cite it in Section 4.2, as well as in two places in the Conclusions where we found its points pertinent to our CMIP7 perspective (see our response to Reviewer 1 Minor comment #19 for the full set of changes).
In Section 4.2:
"Moreover, many 'observed' global-scale products in fact undergo some processing, and in some cases the observed and simulated variables do not represent the same physical quantity despite sharing a name, for example, leaf area index is defined differently across communities (Beadling et al., 2026)"
"Beadling et al. (2026) discuss these issues in detail and recommend best practices for handling observational uncertainty in CMIP-era model evaluation, including the use of multiple independent reference datasets and clearer documentation of uncertainty by data providers."
In the Conclusions, we have additionally noted that machine learning is increasingly applied to the construction of observational reference products as well as to parameterizations, and that realising the evaluation benefits of higher-resolution simulations depends on a parallel expansion of observational capability through new campaigns and the continuation of sustained in-situ measurements.

L1065: "Systematic biases arise from interacting uncertainties and parameterizations governing the exchange of energy, water, and carbon across the Earth system." -> I suggest adding "in Earth System Models and Climate Models" or something after systematic biases to indicate you are referring to biases within climate models, or "CMIP models".

We add "*in coupled climate models*" defined earlier as encompassing Earth System Models and Climate Models."

L1066 – 68: I found this sentence very confusing . . . you can have a bias that is sourced from the ocean (i.e., you would see it in an OMIP run) . . . not everything has to "propagate down through the atmosphere".

We thank the reviewer for this comment and agree that the original sentence implied an incorrect, one-directional pathway in which all biases originate in the atmosphere. We have revised the sentence to
"These biases are rarely confined to a single component. They can originate in any part of the system, for example in atmospheric cloud and radiation processes, ocean mixing and circulation, or land surface processes, and are often amplified through biosphere feedbacks and large-scale climate variability."

L1069: "and simplified land and ocean biogeochemistry, many". **Also simplified in their atmospheric chemistry!** Not all models have interactive ozone for example. And have different aerosol schemes.

We have added this and feel is indeed a very important point and an area often overlooked in CMIP (in this review for example).

L1070: "are persistent across" -> "persist across".

This has been changed.

L1071: "Ocean processes are a significant source of persistent CMIP6 bias because they control long-term heat and carbon uptake (Fig. 6)." -> The reasoning is incomplete as this is written. The ocean processes are not a significant source ***because*** they control long-term heat and carbon uptake. The fact that they play a large role in the planetary heat and carbon budget does not necessarily mean that they will cause a substantial amount of model biases.

You might want to reword to *"Given that the ocean plays a large role in the transient climate response via its control on planetary heat and carbon uptake, biases in ocean processes that persist across model generations can propagate into other realms of the climate system"*

Thanks for making this point and providing an alternative suggestion. We have changed it to: *"Because the ocean strongly influences the transient climate response through its control of planetary heat and carbon uptake, persistent biases in simulated ocean processes can propagate through the climate system, affecting other components and feedbacks."*

L1073: ", which regulate heat and salt." -> This feels incomplete. Which regulates heat and salt transport? Is that what you mean?

We have changed it to: "*the transport of heat and salt.*"

L1076: "thereby affecting **water mass formation**, including deep-water formation and the strength of the global overturning circulation, **which influences the ocean's capacity to sequester heat and carbon**." § See text suggestions. Most of the anthropogenic heat and carbon is going in within water masses that we find in the upper ocean and at intermediate depths (AAIW and mode waters; these are critical for heat and carbon storage).

Thanks for this suggestion, we have taken it on as you have written.

L1077: "AMOC" is not a "climate mode". **Please reword for accuracy.**

Thanks for catching this. We have changed to: "*These biases propagate into large-scale climate modes of variability and circulation features, such as ENSO and the AMOC.*"

L1078: "Systematic biases in sea ice area (SIA) in both the Arctic and Antarctic arise from the spatiotemporal distribution of ice cover, and the response of sea ice to anthropogenic warming." -> **Please reword for accuracy .... These are not what CAUSE the systematic bias.**

We agree this is incorrect as written and have changed to: *"Models exhibit systematic biases in both Arctic and Antarctic sea ice area (SIA), which are evident in the spatiotemporal distribution of ice cover and in the response of sea ice to anthropogenic warming."*

L1083: I suggest transitioning to a new paragraph when you transition to the discussion of land biases.

Perhaps the formatting is unclear but there is a paragraph break here. hopefully this is clear in the final typesetting.

L1093: "permafronst" -> typo.

Thanks for catching this.

L 1128: "affect whether warming stabilizes, continues, or reverses." -> "whether current and future warming ..."

We have added current and future warming as you suggest.

L1130: "long-term committed **climate** warming".

This has been added.

L1133: "or the most relevant diagnostics of the global climate system," -> I would lighten the wording here as this is subjective, one can argue that other diagnostics are more relevant.

We have changed the wording: *"This review provides a model performance overview across various CMIP phases for diagnostics considered to be among the most relevant by the CMIP community, based on those currently considered by the REF."*

L1146 – 47: "Where observations are available, there is not necessarily high confidence surrounding the bias as some diagnostics have uncertainty in their reference datasets." -> this feels vague, can you provide an example?

This is intended as a paragraph opener, where the rest of the paragraph elaborates. We have added concrete examples for clarity:
"Where observations are available, there is not necessarily high confidence surrounding the bias because reference datasets themselves carry uncertainty. For example, global GPP estimates differ substantially between observation-based products, meaning that the magnitude (and in some cases even the sign) of the apparent bias depends on the dataset used for evaluation (Gier et al., 2024). Reference datasets can represent different combinations of observations and modelling assumptions: in situ measurements provide local observations, satellites require retrieval algorithms to convert radiances into physical quantities, while reanalyses combine observations with a numerical model. These can produce substantially different estimates of the same variable: satellite retrievals of cloud properties can differ between sensors and algorithms, precipitation estimates vary between gauge-, satellite- and reanalysis-based products, and soil-moisture estimates differ between reanalyses and observational products."

L1186 – 88: "However, persistent biases in key modes of variability such as ENSO and AMOC are unlikely to be fully resolved, as they emerge from coupled interactions across scales rather than a single model component." -> **AMOC is not a "mode of variability".**

Apologies (again). This has been corrected: *"However, persistent biases in key modes of variability like ENSO and large-scale circulation features like AMOC are unlikely to be fully resolved..."*

L1195: "We also anticipate substantial progress in cryosphere and Earth system coupling. Improved sea dynamics and thermodynamics, together with more realistic icesheet-ocean interactions, are expected to reduce biases in sea ice area and freshwater fluxes (Trivedi et al., 2025)." -> It would be really useful for the authors to provide an indication of how many CMIP7 modeling centers are planning to submit simulations with ice-sheet coupling.

In a survey about the sub models used in each coupled model configuration contributing to CMIP5 and 6, a few had the option to include interactive ice sheets, none of them actually used them for the historical runs submitted to CMIP5 and 6. See `https://doi.org/10.5281/zenodo.7813495` for further details. Some models signed up to ISMIP7 but its not clear that all of these are developed in time for CMIP7 or will use interactive ice sheets in the CMIP runs so we cannot answer that question.

L1215: "This review synthesises current understanding of systematic biases across the Earth system," -> I would rephrase to "current understanding of key systematic . . . " as it is not fully comprehensive.

Done.

Section 5: Since this paper is targeting toward end-users who may not be as familiar with CMIP / coupled modeling. It would be **really useful** to end on a positive note here, regardless of systematic biases (and we have improved in key areas!!) climate models are still useful tools . . . . we just need to choose how to use their output appropriately.

We have given a generally positive ending and included more sucess stories throughout the paper:
"This review does not diminish the value of climate models as tools for understanding and exploring the climate system. Successive generations of models have improved the representation of many important processes and aspects of the observed climate, while many other priorities are expected to be addressed in CMIP7. A key strength of the CMIP multimodel framework is the

ability to evaluate multiple models against common reference datasets, and ongoing efforts such as the Rapid Evaluation Framework will further support the standardisation and robustness of bias assessment across models. The value of CMIP therefore lies not in any individual model being a perfect representation of the Earth system, but in using a coordinated ensemble of models alongside observations and process understanding to identify robust signals and quantify uncertainty. The key is therefore to understand where those biases matter to select and interpret model output appropriately for the question and application. Used in this way, CMIP provides a powerful resource for research, assessment, and decision-making."

References

Beadling, R. L., Russell, J. L., Stouffer, R. J., Mazloff, M., Talley, L. D., Goodman, P. J., Sallée, J.-B., Hewitt, H. T., Hyder, P., and Pandde, A.: Representation of Southern Ocean properties across coupled model intercomparison project generations: CMIP3 to CMIP6, Journal of Climate, 33, 6555–6581, 2020.

Beadling, R. L., Swaminathan, R., Beucher, R., Blockley, E., Brands, S., Hassler, B., Hegedűs, D., Hoffman, F. M., Lee, J., Lewis, J., Lu, J., Malinina, E., Medeiros, B., Scoccimarro, E., Tjiputra, J., Turner, B., and Watson-Parris, D.: Observational Data for Next-Generation Climate Model Evaluation: Requirements, Considerations, and Best Practices, Bulletin of the American Meteorological Society, 107, E813–E835, https://doi.org/10.1175/BAMS-D-25-0079.1, 2026.

Bodas-Salcedo, A., Mulcahy, J. P., Andrews, T., Williams, K. D., Ringer, M. A., Field, P. R., and Elsaesser, G. S.: Strong dependence of atmospheric feedbacks on mixed-phase microphysics and aerosol-cloud interactions in HadGEM3, Journal of Advances in Modeling Earth Systems, 11, 1735–1758, https://doi.org/https://doi.org/10.1029/2019MS001688, tex.eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001688, 2019.

Bushuk, M., Bonan, D. B., Griffies, S. M., Gregory, W., Zhang, Y., Hurlin, B., Chen, Y.-T., Rackow, T., and Goessling, H. F.: Historical and Projected Antarctic Sea Ice Trends Across High-Resolution Coupled Model Hierarchies, Geophysical Research Letters, 53, e2026GL124 043, https://doi.org/10.1029/2026GL124043, 2026.

Forster, P. M., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M. D., Watanabe, M., Wild, M., and Zhang, H.: The earth's energy budget, climate feedbacks, and climate sensitivity, in: Climate change 2021: The physical science basis. Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/9781009157896.009, 2021.

Gier, B. K., Schlund, M., Friedlingstein, P., Jones, C. D., Jones, C., Zaehle, S., and Eyring, V.: Representation of the terrestrial carbon cycle in CMIP6, Biogeosciences, 21, 5321–5360, https://doi.org/10.5194/bg-21-5321-2024, 2024.

Gulev, S. K., Thorne, P. W., Ahn, J., Dentener, F. J., Domingues, C. M., Gerland, S., Gong, D., Kaufman, D. S., Nnamchi, H. C., Quaas, J., Rivera, J. A., Sathyendranath, S., Smith, S. L., Trewin, B., von Schuckmann, K., Vose, R. S., Allan, R., Collins, B., Turner, A., and Hawkins, E.: Changing state of the climate system, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental

Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M. I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J. B. R., Maycock, T. K., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., pp. 287–422, Cambridge University Press, Cambridge, UK, ISBN 978-1-009-15789-6, https://centaur.reading.ac.uk/101849/, 2021.

Hoffman, F. M., Hassler, B., Swaminathan, R., Lewis, J., Andela, B., Collier, N., Hegedűs, D., Lee, J., Pascoe, C., Pflüger, M., Stockhause, M., Ullrich, P., Xu, M., Bock, L., Chun, F., Gier, B. K., Kelley, D. I., Lauer, A., Lenhardt, J., Schlund, M., Sreeush, M. G., Weigel, K., Blockley, E., Beadling, R., Beucher, R., Dugassa, D. D., Lembo, V., Lu, J., Brands, S., Tjiputra, J., Malinina, E., Medeiros, B., Scoccimarro, E., Walton, J., Kershaw, P., Lanfer Marquez, A., Roberts, M. J., O'Rourke, E., Dingley, B., Turner, B., Hewitt, H., and Dunne, J. P.: Rapid Evaluation Framework for the CMIP7 Assessment Fast Track, Geoscientific Model Development, 19, 7415–7455, https://doi.org/10.5194/gmd-19-7415-2026, 2026.

Holmes, C. R., Bracegirdle, T. J., Holland, P. R., Stroeve, J., and Wilkinson, J.: Brief communication: New perspectives on the skill of modelled sea ice trends in light of recent Antarctic sea ice loss, The Cryosphere, 18, 5641–5652, https://doi.org/10.5194/tc-18-5641-2024, 2024.

Hyder, P., Edwards, J., Allan, R., Hewitt, H., Bracegirdle, T., Gregory, J., Wood, R., Meijers, A., Mulcahy, J., Field, P., Furtado, K., Bodas-Salcedo, A., Williams, K., Copsey, D., Josey, S., Liu, C., Roberts, C., Sanchez, C., Ridley, J., Thorpe, L., Hardiman, S., Mayer, M., Berry, D., and Belcher, S.: Critical Southern Ocean climate model biases traced to atmospheric model cloud errors, Nature Communications, 9, https://doi.org/10.1038/s41467-018-05634-2, 2018.

Kaufman, Z., Wilson, E., Purich, A., Beadling, R., and Li, Y.: The Impact of Underestimated Southern Ocean Freshening on Simulated Historical Sea Surface Temperature Trends, Geophysical Research Letters, 52, e2024GL112 639, https://doi.org/10.1029/2024GL112639, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2024GL112639, 2025.

Kim, K., Lee, M.-I., Scaife, A. A., and Smith, D. M.: ENSO phase transition enables prediction of winter North Atlantic Oscillation one year ahead, Nature Communications, 17, 2588, https://doi.org/10.1038/s41467-026-70646-2, 2026.

Lee, J., Planton, Y. Y., Gleckler, P. J., Sperber, K. R., Guilyardi, E., Wittenberg, A. T., McPhaden, M. J., and Pallotta, G.: Robust Evaluation of ENSO in Climate Models: How Many Ensemble Members Are Needed?, Geophysical Research Letters, 48, e2021GL095 041, https://doi.org/10.1029/2021GL095041, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2021GL095041, 2021.

Luo, F., Ying, J., Liu, T., and Chen, D.: Origins of Southern Ocean warm sea surface temperature bias in CMIP6 models, npj Climate and Atmospheric Science, 6, 127, https://doi.org/10.1038/s41612-023-00456-6, 2023.

Miao, J., Wu, X., Jiang, D., Song, X., Wang, T., and Xu, X.: Revisiting Unstable ENSO Teleconnections From a Global Perspective, Journal of Geophysical Research: Atmospheres, 130, e2025JD043 757, https://doi.org/10.1029/2025JD043757, 2025.

Muntjewerf, L., Sacks, W. J., Lofverstrom, M., Fyke, J., Lipscomb, W. H., Ernani da Silva, C., Vizcaino, M., Thayer-Calder, K., Lenaerts, J. T. M., and Sellevold, R.: Description and Demonstration of the Coupled Community Earth System Model v2 – Community Ice Sheet Model v2 (CESM2-CISM2), Journal of Advances in Modeling Earth Systems, 13, e2020MS002 356, https://doi.org/10.1029/2020MS002356, 2021.

Planton, Y. Y., Guilyardi, E., Wittenberg, A. T., Lee, J., Gleckler, P. J., Bayr, T., McGregor, S., McPhaden, M. J., Power, S., Roehrig, R., Vialard, J., and Voldoire, A.: Evaluating Climate Models with the CLIVAR 2020 ENSO Metrics Package, Bulletin of the American Meteorological Society, 102, E193–E217, https://doi.org/10.1175/BAMS-D-19-0337.1, 2021.

Rackow, T., Danilov, S., Goessling, H. F., Hellmer, H. H., Sein, D. V., Semmler, T., Sidorenko, D., and Jung, T.: Delayed Antarctic sea-ice decline in high-resolution climate change simulations, Nature Communications, 13, 637, https://doi.org/10.1038/s41467-022-28259-y, 2022.

Roach, L. A. and Polvani, L. M.: Persistent Antarctic Sea Ice Biases in CMIP6 Models in spite of the Recent Decadelong Sharp Decline, Journal of Climate, 39, 1669–1679, https://doi.org/10.1175/JCLI-D-25-0477.1, 2026.

Scaife, A. A., Dunstone, N., Hardiman, S., Ineson, S., Li, C., Lu, R., Pang, B., Klein-Tank, A., Smith, D., Van Niekerk, A., Renwick, J., and Williams, N.: ENSO affects the North Atlantic Oscillation 1 year later, Science, 386, 82–86, https://doi.org/10.1126/science.adk4671, 2024.

Smith, R. S., Mathiot, P., Siahaan, A., Lee, V., Cornford, S. L., Gregory, J. M., Payne, A. J., Jenkins, A., Holland, P. R., Ridley, J. K., and Jones, C. G.: Coupling the U.K. Earth System Model to Dynamic Models of the Greenland and Antarctic Ice Sheets, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 520, https://doi.org/10.1029/2021MS002520, 2021.

Sohail, T., Irving, D. B., Zika, J. D., Holmes, R. M., and Church, J. A.: Fifty year trends in global ocean heat content traced to surface heat fluxes in the sub-polar ocean, Geophysical Research Letters, 48, e2020GL091 439, https://doi.org/https://doi.org/10.1029/2020GL091439, tex.eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020GL091439, 2021.

Song, Z., Jing, Z., Yuan, M., Li, Z., Wang, H., Song, F., Yang, H., Du, T., and Wu, L.: Reduced Warm Bias in the Surface Southern Ocean by Improved Parameterization of Wind-Driven Vertical Mixing in an Eddy-Resolving Coupled Global Climate Model, Geophysical Research Letters, 52, e2025GL115 504, https://doi.org/10.1029/2025GL115504, 2025.

Spira, T., Dufour, C. O., Frenger, I., Richaud, B., Bennett, M. G., Bitz, C. M., Bracegirdle, T. J., Castellani, G., Challet, F., Chidichimo, M. P., Cocetta, F., Diamond, R., Feba, F., Gülk, B., Guarino, M.-V., Gupta, M., Haumann, F. A., Heil, P., Hoffmann-Abdi, K., Iovino, D., Klocker, A., Lique, C., Martin, T., Massonnet, F., Mezzina, B., Noro, M., Oglethorpe, K., Perez, F. A., S, F. T., Silvano, A., Smith, I. J., and Solomon, A.: No longer polar opposites: the Arctic and the Antarctic are converging towards a seasonally dominated sea-ice regime, Bulletin of the American Meteorological Society, -1, https://doi.org/10.1175/BAMS-D-25-0340.1, 2026.

Stouffer, R., Eyring, V., Meehl, G., Bony, S., Senior, C., Stevens, B., and Taylor, K.: CMIP5 scientific gaps and recommendations for CMIP6, Bulletin of the American Meteorological Society, 98, 95–105, https://doi.org/10.1175/BAMS-D-15-00013.1, 2017.

Sutton, R. T.: Climate science needs to take risk assessment much more seriously, Bulletin of the American Meteorological Society, 100, 1637 – 1642, https://doi.org/10.1175/BAMS-D-18-0280.1, 2019.

Takano, Y., Ilyina, T., Tjiputra, J., Eddebbar, Y., Berthet, S., Bopp, L., Buitenhuis, E., Butenschön, M., Christian, J. R., Dunne, J. P., Gröger, M., Hayashida, H., Hieronymus, J., Koenigk, T., Krasting, J. P., Long, M. C., Lovato, T., Nakano, H., Palmieri, J., Schwinger, J., Séférian, R., Suntharalingam, P., Tatebe, H., Tsujino, H., Urakawa, S., Watanabe, M., and Yool, A.: Simulations of ocean deoxygenation in the historical era: insights from forced and coupled models, Frontiers in Marine Science, 10, 1139 917, https://doi.org/10.3389/fmars.2023.1139917, 2023.

Wang, C., Zhang, L., Lee, S.-K., Wu, L., and Mechoso, C. R.: A global perspective on CMIP5 climate model biases, Nature Climate Change, 4, 201–205, 2014.

Wunderling, N., von der Heydt, A. S., Aksenov, Y., Barker, S., Bastiaansen, R., Brovkin, V., Brunetti, M., Couplet, V., Kleinen, T., Lear, C. H., Lohmann, J., Roman-Cuesta, R. M., Sinet, S., Swingedouw, D., Winkelmann, R., Anand, P., Barichivich, J., Bathiany, S., Baudena, M., Bruun, J. T., Chiessi, C. M., Coxall, H. K., Docquier, D., Donges, J. F., Falkena, S. K. J., Klose, A. K., Obura, D., Rocha, J., Rynders, S., Steinert, N. J., and Willeit, M.: Climate tipping point interactions and cascades: a review, Earth System Dynamics, 15, 41–74, https://doi.org/10.5194/esd-15-41-2024, 2024.

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., and Taylor, K. E.: Causes of Higher Climate Sensitivity in CMIP6 Models, Geophysical Research Letters, 47, e2019GL085 782, https://doi.org/10.1029/2019GL085782, 2020.