Referee Report

Manuscript:
Assessment of the Met Office Global Coupled model version 5 (GC5) configurations

Authors:
P. Xavier and co-authors

Journal:
Geoscientific Model Development
(EGUsphere preprint egusphere-2026-4281)

Recommendation:
Major revisions

Summary of the manuscript

The paper documents the Met Office Global Coupled model version 5 (GC5), formed from the GAL9 atmosphere/land and GOSI9 ocean/sea-ice configurations, and assesses it against its predecessor GC4, against ERA5, and against a range of observational products. The assessment covers global mean state, the hydrological cycle, coupled NWP trials, large-scale variability (ENSO, IOD, storm tracks, blocking, tropical cyclones), and four regional foci (South Asian monsoon, East Asia, Maritime Continent, Africa). The headline result is broad improvement in the global mean state and in NWP scores, with a smaller set of named degradations.

General assessment

Configuration-documentation papers of this kind are useful and GMD is the right venue. The GC5 development represents a large body of work, the breadth of the assessment is impressive, and several of the individual results are interesting in their own right, particularly the improved African easterly jet and associated MCS propagation, the improved Maritime Continent precipitation diurnal cycle, the Antarctic sea ice improvement, and the recovery of tropical cyclone intensity. The paper deserves to be published.

However, the manuscript in its present form is not ready. My concerns fall into four groups, in decreasing order of seriousness:

1. *Nothing is quantified and nothing is tested*. The paper is written almost entirely in the language of "improved," "degraded," "slightly better," and "much improved." A reader cannot tell whether a stated improvement is 2% or 40%, nor whether it exceeds the internal variability of a single 50-year realization. This is the central weakness.

2. *Causal attribution is asserted rather than demonstrated*. Thirty tickets are applied simultaneously and only the net difference is shown, yet the text repeatedly attributes specific changes to specific tickets.

3. *A freshwater conservation error is disclosed only in the Summary*, with no indication of which results in the paper predate the fix.

4. *The manuscript contains a substantial number of internal inconsistencies* between text, figures, and captions, several of which are outright errors (see Specific and Technical comments). At least one figure appears to be mislabeled in a way that changes its scientific meaning.

Major comments

1. Quantification and statistical significance (throughout)

The assessment rests on one 50-year realization of GC5 and one of GC4 (lines 92 to 95), with no ensemble. Differences between two such runs contain an unknown contribution from internal variability. Nowhere in the paper is this addressed.

- Figure 1 is the paper's summary diagnostic and carries no uncertainty information at all. The grey shading is presumably an observational or internal-variability range but it is never defined, and the green/amber/red color coding of the dots depends entirely on it (caption to Fig. 1). Please state explicitly how the "observed range" is constructed, and add an estimate of the sampling uncertainty on each normalized RMSE ratio. Without this, statements like "most tropospheric metrics have improved rather than degraded" (line 134) are unsupported.
- Please add significance testing, or at minimum a bootstrap or block-resampled confidence interval, to the map differences in Figs. 2, 3, 6, 12, 19 and 20. Figure 17 already uses stippling at 95%; extend this practice.
- The AMOC change (18 ± 1 Sv to 16 ± 1 Sv, line 161) is the only place a ± appears in the text. Define what the ± represents. If it is the interannual standard deviation, then the difference of the means is not obviously significant and should not be presented as an established reduction.
- The NWP results rest on two seasons in 2021 and 2022 (line 230). State the number of forecast cases and whether the scorecard triangles in Figs. 11 and 22 encode significance. "Significant improvements ... at all lead times" (abstract, line 15) is a strong claim from a two-season trial.
- Recommendation: run at least a small ensemble for the key metrics, or state clearly and prominently that internal variability is unquantified and that individual regional differences should be read as indicative.

2. Attribution of changes to individual developments

Thirty tickets (26 development, 4 tuning) were applied together and the paper shows only the net GC5 minus GC4 difference. Yet the text repeatedly attributes effects to named changes:

- Line 158: the shortwave contrast change is "via low and mid-level clouds and due to the changes in cloud erosion and Bimodal initiation changes."
- Line 117: increased Sahel rainfall is "apparently linked to a northward movement of the ITCZ ... which may partly result from increased sensible heating in northern Africa following the albedo changes."
- Line 361: TC weakening is attributed to "re-tuning of convection momentum transport."
- Lines 179 to 181: the Antarctic sea ice increase "is attributed to improvements in the Southern Ocean sea surface temperatures."

None of these is supported by a single-ticket sensitivity experiment. Either present such experiments (even short atmosphere-only ones would help), or reword each of these as an explicit hypothesis. As written, the paper reads as if the causal chain is established when it is not.

3. Tuning versus process improvement, and compensating errors

Four of the thirty tickets are tuning (line 73), and the Southern Ocean result explicitly derives from a targeted "Southern Ocean package" (abstract, line 8; lines 84 to 86). The paper does not distinguish improvements arising from better process representation from improvements arising from tuning against the same observations used for evaluation.

The manuscript shows it understands this issue: lines 145 to 147 invoke "well-documented compensating errors on model tuning (Lee et al., 2016)" to explain why cumulative aerosol properties are better simulated than individual species. That insight is correct and important, and it is then dropped. Please generalize it. In particular:

- Mark the four tuning tickets explicitly in Table A1.
- Discuss which evaluation metrics are effectively not independent of the tuning targets.
- Section 6.2 and the abstract note that the N96ORCA1 configuration has a cold bias in global mean temperature requiring further work. A tuning that works at N216 and fails at N96 is direct evidence of resolution-dependent compensating errors. This deserves discussion in the body, not deferral to a future paper.

4. The N96ORCA1 cold bias is deferred but is material

Abstract lines 21 to 25 and lines 492 to 495 state that the low-resolution configuration, which is the basis for UKESM, has a cold bias in global mean temperature in the pre-industrial control that requires continued tuning. This is presented as a forward pointer, but a reader evaluating GC5 needs to know its magnitude. Please state the size of the bias in kelvin and the length of the

run over which it is diagnosed. A configuration that is well tuned at one resolution and cold at another is a significant caveat on the whole assessment, and burying it costs the paper credibility that a frank statement would earn back.

5. The freshwater conservation problem (line 491)

The Summary states: "Freshwater conservation issues were identified and substantially corrected at the post-assessment phase, thereby reducing ocean-ice volume drift toward acceptable levels."

This appears nowhere in the body of the paper. It raises questions the reader cannot answer:

- What was the error, and what was its magnitude in Sv or mm/yr of equivalent sea level?
- Over what period did it act, and what was the resulting ocean and sea-ice volume drift?
- Critically: were the GC5 results shown in this paper produced before or after the correction? If before, which figures are affected and by how much? The Southern Ocean, Antarctic sea ice, AMOC and subsurface temperature results are all plausibly sensitive to a freshwater budget error.
- What does "acceptable levels" mean quantitatively?

This needs a dedicated subsection in the body with numbers. As it stands, a single sentence in the Summary undermines confidence in the ocean and sea-ice results, which is the opposite of what the authors presumably intend.

6. Over-reliance on "not shown" and on external references

I count more than fifteen "(not shown)" attributions, several of them carrying substantial weight:

- MJO propagation and amplitude (line 452). This is listed in the abstract (line 17) and Summary (line 486) as one of the model's principal degradations, and there is no figure whatsoever. The MJO is central to the Maritime Continent section and to sub-seasonal prediction, which is one of the stated applications of GC5. Please add a figure (Wheeler-Kiladis or lag-correlation propagation diagram, plus an amplitude metric).
- Section 3.1 on ENSO and the IOD refers the reader to figures in Li et al. (2023) five times and shows no ENSO diagnostic of its own. A GC5 assessment paper should stand on its own for a mode of variability this important. Please include at least the Niño 3.4 variability and the SSTA standard deviation pattern.
- Other load-bearing "not shown" items: African easterly wave statistics (line 460), Arctic sea ice concentration biases (line 175), upper-level jet climatology and variance (line 427), QBO behavior (lines 197 to 201), latent heat changes (line 114), near-surface salinity (line 160), and JJA storm tracks (line 331).

Please promote the MJO and ENSO material to figures and prune the rest.

7. Land carbon: the GPP bias is understated

Lines 121 to 123: global GPP is "around 85 GT C/yr in GC5 compared to ~120 GT C/yr in the observations," described as "a problem since GC3." A persistent 30% low bias in gross primary productivity, in a configuration that is to become the basis for an Earth system model, warrants more than one sentence and no figure. Please show the spatial pattern, state whether GC5 is better or worse than GC4 in the global mean, and comment on the implications for UKESM.

8. Ocean circulation deficiencies are framed too positively

- ACC transport (lines 166 to 168): 80 Sv in GC4 to 100 Sv in GC5, against an observational estimate of 170 Sv (Donohue et al., 2016). This is described as "a significant improvement but still low." It remains roughly 40% too weak. Please reframe: the direction of change is right, the magnitude of the remaining error is large, and the consequences for Southern Ocean heat and carbon uptake should be noted.
- AMOC (lines 161 to 163): reduced from 18 to 16 Sv and now "slightly low." Given that the Southern Ocean package is the flagship ocean improvement, a weakening of the AMOC in the same model version deserves more than two sentences. Is the reduction consistent across the run, or a drift?

9. Near-surface wind speed: the net effect is not assessed

Lines 130 to 134 state that near-surface wind speed increases globally and that "overall, this is beneficial, as GC4 mostly underestimated" it. The same paragraph and later text then attribute to that increase: degradation where winds were already excessive (line 131), a roughly 10% increase in natural aerosol emission that partly offsets the intended aerosol reduction (line 143), a possible worsening of the Indian Ocean 200 to 300 m subsurface warm bias (line 133), and increased 10 m wind RMSE in NWP (line 240). Please give the net assessment honestly rather than leading with "beneficial."

10. Observational uncertainty

GPCP, GPM-IMERG, CERES-EBAF, HadISST, ESA CCI, PIOMAS, AERONET and ERA5 are all treated as truth. In several places the model-observation difference is comparable to the spread among observational products, notably tropical precipitation and Southern Ocean SST. Please add a short paragraph on observational uncertainty, and where feasible show it (for example, GPCP versus IMERG in Fig. 2a). Note also that Fig. 9 compares GC5 for 1989 to 2008 with GPM observations for 2001 to 2019: different periods, which should at least be flagged.

11. A table of simulations is needed

The paper uses N96ORCA1, N216ORCA025, N320ORCA025, N640ORCA025, N1280 case studies, coupled NWP trials at 40/20/10 km deterministic and 60/40/20 km ensemble,

MOGREPS-UK trials, atmosphere-only GAL9 and GA8 runs, DevGAL9GO6/GLX near-real-time trials, and four N96ORCA1 historical simulations plus a piControl. These appear scattered through the text, sometimes inconsistently (see Technical comments). A single table listing every simulation used, with resolution, ocean configuration, forcing, run length, calendar period, number of members, and the sections and figures in which it appears, would be the single most valuable addition to this paper. I would place it in Section 1.

12. "Switching off deep convection" deserves a proper description

Ticket 557 (given as 577 in the text, line 45; see Technical comments) removes the separate deep convection scheme entirely and represents surface-driven deep convection through the modified mid-level scheme. This is the most structurally significant change in GAL9 and it receives five lines. Please expand: what closure applies, how CAPE removal behaves, what happens to the diurnal cycle over land, and how the scheme behaves at the range of resolutions used here. The improved Maritime Continent diurnal cycle (Fig. 20), the improved AEW and MCS propagation (Fig. 21), and the shift toward moderate precipitation intensities (Figs. 9 and 10) are all plausibly consequences of this single change, and the paper never makes that connection. Doing so would considerably strengthen the scientific content.

13. Model drift and conservation diagnostics

For a 60-year coupled present-day simulation, please report the TOA net radiative imbalance in W/m², the global mean surface temperature trend, and the ocean heat content drift, for both GC4 and GC5. Line 154 says there is "no notable change in global mean radiative balance," which is not a substitute for the numbers. Given the freshwater issue noted above, global salinity and sea-ice volume drift should also be shown.

14. Code availability

The Code and Data Availability statement notes that UM and JULES source code cannot be provided due to intellectual property restrictions, with a copy made available to reviewers. The editor should confirm this meets GMD's code and data policy. I would also ask the authors to confirm that the Zenodo archive (doi:10.5281/zenodo.21353930) is live and contains the analysis code for every figure, including those adapted from internal Met Office material (see the comment on Fig. 14).

Specific comments (by line number)

Abstract

- L3: "more than 26 science tickets" conflicts with L73, "26 development tickets and 4 model tuning tickets." Use one number consistently.
- L15 to 16: "significant improvements ... at all lead times were achieved." State the magnitude and the significance test.
- L19 to 20: "will require minimal adjustments." This is asserted without support and is contradicted in tone by Section 6.1, which reports that GC5 degrades MOGREPS-UK for surface temperature, 10 m wind and cloud, with the degradation growing with lead time. Either soften the abstract or strengthen the evidence.
- L21 to 25: quantify the N96ORCA1 cold bias (see Major comment 4).
- The abstract lists improvements in detail and degradations in a single compressed sentence. Balance these.

Section 1

- L27 to 29: give the date of the intended GC5 operational implementation if known, and state GC5's role in CMIP7 as well as CMIP6 if applicable, since CMIP6 is now well past its main phase.
- L45 to 49: see Major comment 12.
- L50 to 53 (FCAPE): "the CAPE closure responds in a more physical manner to inversions, etc." Replace "etc." with the actual cases, and give evidence for "more physical."
- L59 to 62 (Fountain Buster): the description of the "eternal fountain" is clear and welcome, but please give the magnitude of the correction term and where in the world it is most active.
- L88 to 91: "A suite of 24 tropical cyclone case studies" conflicts with L364, "Fourteen N1280 case studies." Also, L89 calls the N640 runs "case studies" while L363 calls them "three-month summer and winter trials." Resolve both.
- L94: constant year-2000 forcings justified by a citation to IPCC (2007). AR4 is an odd and dated authority for this choice. Cite the CMIP6 or AMIP protocol instead.
- L95: "Years 10 to 60 of the simulation are assessed." State the total run length and the spin-up rationale. This also conflicts with the Fig. 7 caption, which describes "the 100-year model run from 1978 to 2078."

Section 2

- L99 to 106: this paragraph is a list of regions where precipitation went up or down, with no figure reference until the last sentence and no indication of magnitude. Consider replacing most of it with a pointer to Fig. 2 and retaining only the results that matter.
- L110 to 114: "Sensible heat has increased on land ... Generally, latent heat has changed relatively little on land, except in India and the Sahel." Give numbers in W/m² and show the fields, or drop the claim.
- L121 to 123: see Major comment 7.
- L126: "the SST bias in the western equatorial Indian Ocean in GC5 is 0.25 to 0.4 K higher than in GC4." This is one of the few quantified statements in the paper. Please apply the same standard elsewhere.

- L134 to 135: "most tropospheric metrics have improved rather than degraded, which is reflected by the overall assessment metric at the bottom of Fig.1." State the value of that overall metric.
- L150 to 151: "these changes bring GC5 into much closer agreement with aerosol size distributions observed in remote regions." Figure 5 shows the GC5 model total peak at roughly 470 cm^{-3} against an observed peak near 200 cm^{-3} at AMAZE, so the model still overestimates by a factor of about 2.4, improved from about 3.7 in GC4. Please state this rather than "much closer agreement."
- L154: "no notable change in global mean radiative balance and their breakdown." Grammar ("their" has no antecedent), and see Major comment 13.
- L155 to 158: the description of the shortwave bias pattern change is hard to follow. Consider restructuring around Fig. 6 panel by panel. Note also that the area-weighted RMS difference against CERES increases from 7.46 in GC4 to 8.51 in GC5 according to the panel annotations in Fig. 6, which the text does not mention. If shortwave cloud forcing error has increased, say so.

Section 2.1

- L197 to 201: the QBO discussion concludes that the period "is highly sensitive to the scaling parameter for launch level fluxes and is thus to some extent tuneable." This is an honest and useful statement. Please give the GC4 and GC5 QBO periods and amplitudes against observations, even in a table.

Section 2.2

- L205: "precipitation and evaporation both remain similar at around 3.02 mm/day." Give the observed value for comparison. L252 to 253 concedes that both are overestimated relative to observations, so the reader should be told the size of the excess here.
- L208 to 210: the text refers to "timestep intensities" but Fig. 9 shows 3-hourly accumulations. Either add the timestep panel or correct the reference.

Section 2.3

- L229 to 234: state the number of forecast cases and the verification methodology.
- L236 to 238: "an increased cool bias in the lower troposphere in GC5 affects tropospheric temperature and geopotential height scores." Quantify.
- L241 to 246: "The data assimilation statistics show a better fit of the background forecast to observations with resulting smaller analysis increments, suggesting a more balanced NWP system using GC5." This is a good result. Show it.

Section 3.1

- L263 to 281: see Major comment 6. This section is essentially a summary of Li et al. (2023).

- L284 to 307: this text is substantially duplicated. The Cai et al. (2011) argument about Rossby waves from the tropical Indian Ocean appears at L285 to 287 and again at L298 to 300. The IOD teleconnection to southern Australia appears at L288 to 294 and again at L301 to 307, in near-identical wording, including the sentence "In both models, the relationship between IOD and winds over the central Indian Ocean basin is too weak compared to that seen with ERA5." Please delete one copy.

Section 3.2

- L318 to 320: RMSE values of 1.37 and 1.26 are given without units or an uncertainty estimate. State the units (presumably cyclones per month per unit area) and whether the difference is significant.

Section 3.3

- L352 to 355: "there has been relatively little change in the blocking frequency statistics from GC4 to GC5." This is a null result and is fine, but Fig. 13c and 13f suggest GC5 moves closer to ERA in the ATL domain and away from it in the PAC domain. Please state the net.

Section 3.4

- L357 to 362: "The operational implementation of GC4 in May 2022 introduced a marked weakening of tropical cyclones." Some readers will not know the operational history. A sentence of context would help.
- L369 to 372: the DevGAL9GO6/GLX near-real-time trial is described as "for internal use only." Clarify whether the results shown in Fig. 14 are reproducible from the archived data.
- L373 to 377: the Hinnamnor and Ian examples are two cases. Please state that they are illustrative and not evidence of systematic skill.

Section 4.1

- L392 to 397: the moisture budget analysis following Keane et al. (2021) is one of the more substantive pieces of process analysis in the paper and is well done. Consider expanding it, since it is the kind of content that distinguishes an assessment paper from a scorecard.
- L399 to 405: "The bias in July and, particularly, August is actually smaller in the coupled simulations than the atmosphere-only simulations, even though the latter are driven by observed SSTs." This is an interesting result that deserves emphasis as evidence of compensating errors. Connect it to Major comment 3.

Section 4.2

- L420 to 431: this subsection is thin, relies on "not shown" twice, and its conclusions ("GC5 seems to have improved to represent zonal (CGT-like) and meridional (PJ-like) waves but even

GC5 tends to overestimate amplitudes") are hedged to the point of not saying much. Either strengthen it with a quantitative metric or shorten it considerably.

Section 4.3

- L444: "The land precipitation bias in DJF is significantly improved in GC5 (Fig.2e)." Figure 2e shows a GC5 minus GC4 difference, not a bias against observations. Point to a bias panel, or reword.
- L448 to 450: "It is interesting to note the relatively low correlations along the coastal grid points and requires further investigation." Coastal diurnal cycle errors in this region are a well-known and important model deficiency. A sentence on the likely cause (land-sea breeze resolution, convective triggering) would be more useful than deferral.
- L452: see Major comment 6 on the MJO.

Section 4.4

- L455 to 470: this is the strongest section of the paper. The AEJ structure improvement, its thermal wind interpretation via reduced regional temperature bias and a better Saharan heat low, and the resulting MCS propagation shown in Fig. 21 form a coherent physical argument. Please show the AEJ fields rather than saying "not shown" (L466), and show the AEW curvature vorticity statistics (L460).

Section 5

- The Summary introduces two pieces of information that do not appear in the body: the freshwater conservation correction (L491) and the N96ORCA1 cold bias detail (L493). Move both to the body. A summary should not contain new results.
- L494 to 496: "Despite these outstanding issues, GC5 represents a significant step forward." Given the Indian Ocean degradation, the MJO degradation, the MOGREPS-UK degradation, the N96 cold bias, and the freshwater issue, this conclusion should be argued rather than stated.

Section 6.1

- L505 to 512: GC5 degrades the operational UK convective-scale ensemble for surface temperature, 10 m wind and cloud, with the degradation growing with lead time, and reduces ensemble spread while increasing error. This is a material operational finding that is reported candidly, which I appreciate. It should be reflected in the abstract in stronger terms than "will require minimal adjustments."

Section 6.2

- L513 to 523: four historical members plus a piControl are mentioned but no results are shown. Either show something or reduce this to a forward pointer.

Figures

Figure quality is a serious weakness throughout. Several figures appear to be lifted from internal presentations, several captions do not match their panels, and several are illegible at print size.

- Fig. 1: Axis labels are unreadable at printed size; there are roughly 40 metric names in a few centimeters. Split into two or three panels grouped by variable. The legend contains Rose suite identifiers (u-bs902, u-co779) which are meaningless to external readers; move them to a table or the data availability statement. Define the grey shading. Define "the observed range" used for the color coding. Give the value of the "Overall assessment" metric in the text.
- Fig. 2: Caption should state that panel (a) is against GPCP (the panel title says so, the caption does not). Panels (c) to (f) use different map projections and aspect ratios from (a) and (b), which is fine, but the colorbar for (c) to (f) is not shown separately; the caption says "the same contour levels and colour scale are used in all panels," which cannot be right given that (a) and (b) are biases against GPCP and a model difference respectively. Please clarify.
- Fig. 3: Caption reads "(c) GC4; (b) GC5" for the bottom row. The second should be (d).
- Fig. 5: The y-axis ranges differ between the GC4 and GC5 columns (1000+ versus 500 in the top row, 1500 versus 800 in the bottom row). The caption acknowledges this, but it makes the visual comparison actively misleading: the GC5 panels look similar to the GC4 panels while representing roughly half the concentration. Use common axes.
- Fig. 6: Panels are labeled b), c), d) with no panel a). The caption refers to "left," "middle," "right." Relabel consistently. Panel titles again contain Rose suite IDs. The area-weighted RMS differences annotated on panels c and d (7.46 for GC4, 8.51 for GC5) show a degradation that the text does not discuss.
- Fig. 7: Three problems. (i) The single left panel is labeled "c)" with no a) or b). (ii) The right-hand panel titles read "GC3minus HadISST" and "GC4minus HadISST," while the caption says the panels are GC5 (left) and GC4 (right). Either the panel title has a typo for GC5 or the wrong figure has been included. This must be resolved. (iii) The caption describes "the 100-year model run from 1978 to 2078," which is inconsistent with the constant year-2000 forcing and the years 10 to 60 assessment period stated at L94 to 95. Explain.
- Fig. 9: Shows only GC5 and GPM, but the surrounding text compares GC5 with GC4. Add the GC4 panels or restrict the text. Model period (1989 to 2008) and observation period (2001 to 2019) differ; state why and whether it matters. Panel titles are too small.
- Fig. 10: The caption contains the sentence "Similar plots for N96 resolutions are also shown." twice. The shaded "Light/Medium/Heavy/Very heavy" bands are not defined.
- Fig. 11: Illegible at print size. Roughly 30 rows by 30 columns by three panels of small triangles. State whether the triangles encode statistical significance and what the size scaling means quantitatively ("max = 20" is unexplained).
- Fig. 12: The caption says panels (c) and (d) are JJA for the southern hemisphere, but all four panel titles read DJF, and the text at L325 explicitly discusses Fig. 12c and 12d "in DJF," consistent with L317 to 318, which says only DJF is shown. The caption is wrong and must be corrected. Also state the units of the RMSE values.

- Fig. 13: The caption is garbled: "For ERA/IV, the asterisk is the mean estimate from ERA whereas (which is also denoted by the blue dashed lines)." The caption describes the boxplots as being "over the ATL domain" but both ATL and PAC are plotted in (c) and (f). Panels (a), (b), (d), (e) are unlabeled as to which is GC4 and which GC5 within the figure itself.
- Fig. 14: This appears to be a screenshot from an internal Met Office presentation. It carries a Met Office logo, an embedded free-text box, a legend in a different font from every other figure in the paper, and a title referring to "MOGM v. GLX" without definition in the caption. Please redraw both panels in the paper's style, define MOGM and GLX in the caption, and confirm the underlying data are in the Zenodo archive.
- Fig. 15: Six panels with a nine-entry legend of overlapping line styles. Very hard to read. Consider splitting, or reducing the number of curves.
- Fig. 16: Caption reads "along with the changes between the GC5 and GC5 (top right)." Should be GC5 and GC4. Also, the caption lists the panels in the order "ERA5 reanalyses, GC5 and GC4," but the panel layout is ERA5 (top left), anomalies (top right), GC4 (bottom left), GC5 (bottom right). Relabel the panels (a) to (d) and describe them in order.
- Fig. 20: The two top panels have different y-axis ranges and the axis label appears only on the left. The bottom-panel colorbar runs from -0.8 to 0.8 but the caption discusses correlations of -1.0.
- Fig. 21: Hovmöller panels have no colorbar label beyond the numbers; state the units in the panel as well as the caption. The observation panel is 2008 while the model panels are "an arbitrary year," which the caption explains well.
- Fig. 22: The scorecard at the top has no identification of which configurations are being compared within the figure. "max = 20" and "3 grid lengths" are unexplained.

Technical corrections

- L45: ticket "577" for switching off deep convection. Table A1 lists this as gmed557. Correct one of them.
- L79: "Van Genuchten et al. (1991)" is cited for soil pedotransfer functions, and the reference list gives the RETC code report (van Genuchten, Leij and Yates, 1991). The pedotransfer function form is normally attributed to van Genuchten (1980), *Soil Sci. Soc. Am. J.*, 44, 892 to 898. Check which is intended.
- L79: "Brooks, 1965" should be Brooks and Corey (1964), *Hydraulic Properties of Porous Media*, Hydrology Papers No. 3, Colorado State University. Both the year and the missing second author need correcting.
- L86: "SI³" appears variously as SI3 and SI^3 through the paper. Standardize.
- L173: "excessive ice north of Iceland and too little ice in the Labrador Sea and Gulf of St. Lawrence."
- L181: "There is a minor reduction in sea ice the Weddell Sea" is missing a preposition ("in the Weddell Sea").
- L183: "(Ridley et al. (2022))" has doubled parentheses. Occurs elsewhere too, e.g. L310 "(see Hoskins and Hodges (2002, 2005); Priestley et al. (2020))."

- L222: "GC5 models continue to exaggerate" should be "GC5 continues to exaggerate," since only one model is discussed.
- L246: "does not negatively impact the CRPS skill measure."
- L268: "SSTa" is used for SST anomalies without definition.
- L440: "A detailed description of the impact of cold surges is given in Lim et al. (2017) and Xavier et al. (2020)" is missing a comma before "and" and the sentence at L439 to 440 duplicates part of L434 to 437.
- L491: "thereby reducing ocean-ice volume drift toward acceptable levels" mixes American and British spelling conventions with the rest of the manuscript ("toward" versus "towards" used at L219). Standardize.
- Reference list: Walters et al. is given as "Geosci. Model Dev. Discuss., p. in press., 2018." The published version is *Geosci. Model Dev.*, 12, 1909 to 1963, 2019, doi:10.5194/gmd-12-1909-2019. Update.
- Reference list: Willett and Others (2026) has no volume, pages or DOI, and the author list is given as "Willett, M. and Others." Since this paper defers most of the GAL9 description to that reference, its status matters. Provide full details or a preprint DOI.
- Reference list: Bush et al. (2024) is cited as "Geoscientific Model Development Discussions, 2024, 1 to 58, 2024." Check whether the final version is now available.
- Reference list: Hardacre et al. (2021) has a DOI but no volume or page numbers.
- Reference list: IOC and IAPSO (2015) entry is malformed: "IOC, S. and IAPSO: The International thermodynamic equation of seawater–2010 ... [includes corrections up to 31st October 2015].., 2015."
- Reference list: Van Genuchten et al. (1991) entry has no publisher or report number.
- Reference list: Xavier et al. (2023) is a Met Office internal Forecasting Research Technical Report. Provide a public link if one exists, since it is the GC4 assessment against which this entire paper is framed.
- Appendix A: Table A1 covers GAL9 tickets only. There is no equivalent table for GOSI9. Please add one, or explain why not.
- Appendix A: mark the four tuning tickets so the reader can distinguish them from development tickets.
- Section 2 is titled "Global Characteristics" but Section 2.3 covers high-resolution NWP trials, which are not a global characteristic. Consider restructuring into mean state, variability, and forecast performance.
- Jargon used without definition on first appearance: MOGM (L~372 figure), GLX (L370, defined loosely), DevGAL9GO6 (L370), RC4 (Fig. 1 legend), WMC (defined at L216, good), AGP (L338), HiRA (Fig. 22 caption).

Conclusion

This is a useful and needed paper describing a large and evidently careful body of model development work, and several individual results (the African easterly jet, the Maritime Continent diurnal cycle, Antarctic sea ice, tropical cyclone intensity) are clearly valuable. But the manuscript currently asks the reader to accept a long list of qualitative claims without

quantification, without uncertainty estimates, and in several places without the figures that would support them. It also contains a number of internal inconsistencies, at least two of which (Fig. 7 panel labels, Fig. 12 caption) must be resolved before a reader can interpret the figures at all, and it discloses a freshwater conservation problem in the Summary without saying how it affects the results presented.

I recommend major revisions. The essential requirements for a revised version are:

1. A table of all simulations used.
2. Quantitative statements with uncertainty estimates in place of "improved" and "degraded," and significance testing on the map differences.
3. Figures for the MJO and for ENSO, and a substantial reduction in "not shown."
4. A full account of the freshwater conservation issue in the body, including which results it affects.
5. Quantification of the N96ORCA1 cold bias.
6. Correction of the figure and caption errors listed above, and removal of the duplicated text on page 17.
7. Reframing of causal attributions as hypotheses unless supported by sensitivity experiments.

I would be happy to review a revised version.