Author response for GMD-2016-161

Dear Dr. Ullrich,

Enclosed please find our response to the Interactive Discussion on manuscript GMD-2016-161, "ASoP (v1.0): A set of methods for analyzing scales of precipitation in general circulation models."

This PDF contains a point-by-point response to each of the reviewers' comments. We have also specified the changes that we have made to the manuscript in response to these comments. The PDF concludes with marked-up version of the manuscript.

The substantial changes to our manuscript are:

- A set of summary metrics for spatial and temporal coherence of precipitation in models and observations, as suggested by the reviewers. A description of these metrics can be found in Section 2.1.4; the values of the metrics for the datasets we analyse are shown in Table 3; the metrics are discussed throughout Section 3 and referenced in Sections 4 and 5.
- A new figure (Fig. 7 in the revised manuscript) showing the spatial and temporal character of precipitation in a model with temporally and spatially persistent precipitation (GEOS5) and a model with temporally and spatially persistent precipitation (MetUM-GA3), as suggested by Reviewer #2.
- Revisions to our discussion of the lack of verifying observations for simulated timestep precipitation data, as suggested by Reviewer #2.
- Revisions to our discussion of the TRMM 3B42 and CMORPH algorithms for estimating precipitation from microwave and infrared imagers, as suggested by Reviewer #2.

There are many sentence-level revisions to our text, as suggested by the reviewers. These are listed in our responses to the reviewers' comments.

Further, we have modified our terminology to refrain from referring to "persistent models" or "intermittent models", in favor of "models with persistent precipitation" or "models with intermittent precipitation." This change was not suggested by the reviewers, but was a decision taken by the co-authors.

Finally, we are pleased to note that the code for our ASoP1 diagnostics will be made publicly available through GitHub, under the terms of the Apache 2.0 license. The link to our GitHub repository can be found in the "Code availability" section of the revised manuscript.

Yours sincerely,

Dr. Nicholas Klingaman
on behalf of co-authors Gill Martin and Aurel Moise

GMD-2016-161
Response to comments from Christian Jakob

We thank the reviewer for providing a positive and encouraging review of our manuscript. In our response below, the reviewer's comments appear in red text; our response to the reviewer's comments appears in black text.

General comments

This is an excellent paper that introduces a number of interesting and useful diagnostics to study the behaviour of precipitation in weather and climate models at various space and time-scales all the way to single grid points and time steps. The paper is very well written and the analyses and arguments are sound. It adds several new ideas for model analysis and evaluation, all of which will add to the arsenal available to the community. I have a few minor comments, which I list below.

We thank the reviewer for these positive comments.

Specific comments

1) The new diagnostics introduce are very nice and worth looking at in the present form. However, as there are many of them, it would be nice if the authors could consider producing a few summary measures that could be presented more easily when comparing many models and/or evaluating them against observations. This would contribute to the growing interest in having "performance metrics", so that changes in models can be more easily assessed and their effects quantified. I am aware that there is no single metric that identifies good or bad models, but by having a collection of them—well beyond this study—I believe the community will be able to better communicate model improvement in the future. For one-dimensional histograms, there are simple statistical techniques one could use, such as the Kolmogorov-Smirnov two sample test, which allows an assessment of the likelihood that two samples are drawn from the same population. This would be especially useful where models are compared to observations. It would be nice to also have a summary measure of the two-dimensional histograms presented here, but that might be more difficult.

We agree with the reviewer and thank the reviewer for this suggestion. Reviewer #2 also proposed that we quantify the persistence and intermittency of precipitation in models and observations. We have combined the two reviewers' suggestions and provide a unified response below, a copy of which appears in our response to Reviewer #2.

We have created summary metrics of spatial and temporal coherence in precipitation, which allow the reader to more easily evaluate models, either against each other or against observations. These metrics are based on the persistence of upper- and lower-quartile precipitation in time (measured from one timestep to the next) or space (measured at neighboring gridpoints). Considered together, the metrics summarise aspects of our two-dimensional histograms, as well as our correlations of precipitation as functions of distance and time, but without relying on the choice of a threshold correlation value or spatial or temporal scale. The metrics are scaled to range from -1 to +1: positive values indicate that persistence is more common than intermittency; negative values indicate that intermittency is more common than persistence. Table 3 of the revised manuscript shows these metrics for all models, as well as for TRMM and CMORPH satellite-derived observations, and for all horizontal grids and temporal frequencies considered in our study (i.e., timestep and 3-hr, native-grid and 5.6°×5.6° averages). The metrics confirm many of the conclusions of our study concerning the effects of averaging in space and time on the spatial and temporal coherence of simulated precipitation features. These quantitative metrics also confirm our qualitative conclusions about the relative levels of spatial and temporal coherence in the models we analysed.

We have added a description of these metrics to Section 2.1.4. The metrics are shown in Table 3. We discuss the metrics throughout the Results section (Section 3) and refer to them in the Discussion (Section 4) and Conclusions (Section 5) sections.

These metrics have helped to demonstrate quantitatively the qualitative conclusions we had drawn from our analysis, so we are very grateful to the reviewers for suggesting them.

2) Page 12, Line 6-7: This is a very strange argument. It's likely you did look at the lagged time correlations for 3-hourly rainfall and found something. Why not just state what you found—no need to show a figure if you can say it in words. Saying that there could be a problem, rather than that there is one, sounds like you have something to hide.

We agree with the reviewer. We have changed the paragraph in question to read: "For model data, lagged correlations of 3-hr precipitation (i.e., as in Fig. 6b but for 3-hr data) over a 36-hr window were dominated by the overly strong and regular diurnal cycle of precipitation in the models, which manifested itself in our diagnostics as a pronounced peak in the correlations at a 24-hr lag (not shown). TRMM and CMORPH displayed a much weaker and broader peak across lags of 18–30 hr, suggesting greater day-to-day variability in the timing of the diurnal maximum in tropical maximum in the satellite-observations than in the models."

We believed that because the lag-correlation diagrams for 3-hr data were dominated by the diurnal cycle in the models, showing the diagrams would not address our original objective of analysing the temporal coherence in the 3-hr precipitation data on sub-daily scales (i.e., lag-1 or lag-2 correlations in the 3-hr data). Instead, it would unduly focus the reader's attention on the errors in the diurnal cycle, which are outside the scope of our study.

3) Page 13, Line 28-29: The atmosphere is not in radiative convective equilibrium at the scale of 600 km over 3 hours. If the models were, that would be completely wrong. To provide evidence, I attach an unpublished figure from my own work, which plots daily averages of atmospheric cooling derived from CERES observations against daily averages of rainfall from GPCP for increasing areas centered on a 1x1 degree gridpoint in the Tropical Western Pacific. If we consider the 90x90 degree area as close to RCE, it is evident that the averaging scale one could argue starts to approach it is about 30-40 degrees (i.e., $(3000\text{-}4000\ \text{km})^2$), far from the 600 km scale speculated about here. In fact, not surprisingly, at those scales, the atmosphere is a far from RCE as it can be, as the heavy rainfall is associated with large cloud fields that reduce the atmospheric radiative cooling substantially. More likely, at the 600 km scale the convection is in balance with dynamical systems at the synoptic scale, which do exist in the tropics, and are perhaps relatively well captured by the models.

We agree with the reviewer and thank the reviewer for this suggestion. We have replaced the sentence in question with: "We hypothesize that these broader scales represent those at which simulated convection is in balance with the synoptic-scale, dynamical systems that produce precipitation, predictions of which should be highly similar among the models in the short, 2-day hindcasts we analysed."

We replaced a similar sentence in the Discussion section with: "This convergence of model behavior may be enhanced by the fact that these data are from short (2-day) forecasts initialized from the same ECMWF analyses, which means that the representation of these dynamical systems are much more similar among models than if the data came from free-running climate simulations."

4) Page 6, Line 7-8 and Figures 2 and 5: The bin below 0.5 * Delta x appears pointless and makes for an awkward plot. I suggest changing the txt to acknowledge the theoretical possibility of the existence of such a bin but then states that in practice it does not exist for this study. Then you can remove the unnecessary and distracting XXX columns from the figures.

We agree with the reviewer. In the revised manuscript, we have removed the bin below $0.5\Delta x$ from our analysis and the column from our revised versions of Figs. 2 and 5.

GMD-2016-161
Response to comments from Anonymous Referee #2

We thank the reviewer for providing a thorough and generally positive review of our manuscript. In our response below, the reviewer's comments appear in red text; our response to the reviewer's comments appears in black text.

General comments

This paper proposes a diagnosis method for precipitation of general circulation models (GCMs) by using a native temporal and spatial grids and discusses dependency of temporal and spatial averages of precipitation. Precipitation behaviors of GCMs have been usually evaluated by climatological mean states. However, this study clearly shows that even if the climatological mean (or 3-hr average) precipitation is almost the same, its temporal and spatial behaviors are very different if analyzed by the native grids and original time step. This aspect of precipitation might affect large scale behaviors and hence must be more focused for the analysis, evaluations, or improvements of GCMs. The methodology is clear, and it implication is sound. Thus, I suggest publication of this study after minor revisions described below. Although the proposed diagnosis will be useful, the authors can go further more. In the text, the authors mention "persistency" or "intermittency" of precipitation. We need to compare many figures (e.g. Fig. 4 vs Fig. 7) to evaluate "persistency" or "intermittency". The authors should consider some quantifications of "persistency" and "intermittency", and show summary of these quantities of the models with difference samplings.

We agree with the reviewer and thank the reviewer for this suggestion. Reviewer #1 (Christian Jakob) also proposed that we quantify the persistence and intermittency of precipitation in models and observations. We have combined the two reviewers' suggestions and provide a unified response below, a copy of which appears in our response to Reviewer #1.

We have created summary metrics of spatial and temporal coherence in precipitation, which allow the reader to more easily evaluate models, either against each other or against observations. These metrics are based on the persistence of upper- and lower-quartile precipitation in time (measured from one timestep to the next) or space (measured at neighboring gridpoints). Considered together, the metrics summarise aspects of our two-dimensional histograms, as well as our correlations of precipitation as functions of distance and time, but without relying on the choice of a threshold correlation value or spatial or temporal scale. The metrics are scaled to range from -1 to +1: positive values indicate that persistence is more common than intermittency; negative values indicate that intermittency is more common than persistence. Table 3 of the revised manuscript shows these metrics for all models, as well as for TRMM and CMORPH satellite-derived observations, and for all horizontal grids and temporal frequencies considered in our study (i.e., timestep and 3-hr, native-grid and 5.6°×5.6° averages). The metrics confirm many of the conclusions of our study concerning the effects of averaging in space and time on the spatial and temporal coherence of simulated precipitation features. These quantitative metrics also confirm our qualitative conclusions about the relative levels of spatial and temporal coherence in the models we analysed.

We have added a description of these metrics to Section 2.1.4. The metrics are shown in Table 3. We discuss the metrics throughout the Results section (Section 3) and refer to them in the Discussion (Section 4) and Conclusions (Section 5) sections.

These metrics have helped to demonstrate quantitatively the qualitative conclusions we had drawn from our analysis, so we are very grateful to the reviewers for suggesting them.

Specific comments

p. 3, L18, "Such diagnostics": It is not clear which "diagnostics" is referred to in this paragraph. Please clarify.

We were referring to diagnostics that estimate the coherence of precipitation in space and time, which we mentioned in the previous sentence. In the revised version of the manuscript, we have replaced "such diagnostics" with "diagnostics of precipitation coherence".

p. 4, L5-6, "Both products are derived from a combination of infrared and microwave sounders and calibrated against gauge data.": The authors should add more information on the difference between TRMM and CMORPH for readers who are not familiar to the details of the products of precipitation. In addition, since TRMM 3B42 is not solely based on the TRMM data, it is not appropriate to call it "TRMM". The authors should make a remark on it if the abbreviation of "TRMM" is to be used.

We disagree with the reviewer that it is inappropriate to call the 3B42 product "TRMM". In fact, the creators of the dataset, the U.S. National Aeronautics and Space Administration (NASA), refer to the product as "TRMM 3B42". Please see their webpage here:

http://disc.sci.gsfc.nasa.gov/precipitation/documentation/TRMM_README/TRMM_3B42_readme.shtml

The Tropical Rainfall Measuring Mission (TRMM) is a project, which includes a satellite that is—confusingly, we agree—also called TRMM. The microwave instrument on the satellite is often called the TRMM Microwave Imager (TMI). Some of the products that TRMM (the project) produces use data exclusively from TRMM (the satellite), while other TRMM (the project) products use data from multiple instruments, including TRMM (the satellite). The 3B42 product falls into the latter category. Many studies, and NASA itself, refer to the 3B42 product as "TRMM 3B42." It is not our place to correct this confusion; rather, we use the commonly accepted nomenclature.

We have added details to Section 2 about the TRMM 3B42 and CMORPH algorithms, including differences in how they produce their merged microwave–infrared precipitation estimates. We have also added a note that states that the TRMM 3B42 product is produced from multiple microwave sounders, not solely TMI. We have also added "3B42" to all figure captions and tables where we use TRMM 3B42 data.

p. 5, L23, "We find the central point in each region and extract the timeseries of precipitation.": I suggest that "find" should be replaced by "define" or an appropriate word.

We have replaced "find" with "select". We hope that this is acceptable.

p. 5, L26, "in Figs. 2b and 2c for CMORPH": These should be "Figs. 2c and 2d".

We agree with the reviewer. We have corrected this error by replacing "in Figs. 2b and 2c" with "in Figs. 2c and 2d".

p. 6, L10, "these computations result in a matrix of correlations with distance and time, as shown in Fig. 2c.": I guess that this is for Figs. 2e and 2f.

We agree with the reviewer. We have corrected this error by replacing "in Fig. 2c" with "in Figs. 2e and 2f".

p. 6, L14, "Fig. 2b": This should be replaced by "Figs. 2c and 2d".

We agree with the reviewer. We have corrected this error by replacing "in Fig. 2b" with "in Figs. 2c and 2d".

p. 6, L16, "For the ranges shown here, the CMORPH 0.25° correlations decline more quickly with time than with space.": It is ambiguous to say which is "more quick" between time and space. Add more explanations.

We thank the reviewer for pointing out this ambiguity. On reflection, this sentence does not contribute anything meaningful to our discussion and is a likely source of confusion. We have removed it from the revised manuscript.

p. 8, L20, "The 1D histogram suggests that MetUM-GA3 oscillates between lighter (< 9 mm day^{-1}) and heavier (> 30 mm day^{-1}) rain rates, with almost no instances of moderate rates (9–30 mm day^{-1}).": This is an interesting behavior of precipitation of the MetUM-GA3. Please consider adding sample figures of time sequence and snapshot distribution of precipitation of MetUM-GA3.

We agree with the reviewer. We have added a figure to our revised manuscript (Fig. 7) that compares MetUM-GA3 to GEOS5. GEOS5 produces persistent rainfall and has a timestep and horizontal resolution similar to MetUM-GA3. In Fig. 7, we compare timeseries of precipitation at an example gridpoint in the Indian Ocean (0°, 90°E) for an example 48-hour forecast from our datasets (4 November 2010); we also compare snapshots of instantaneous precipitation rates for an example timestep (20:00 UTC on 4 November 2010). These figures confirm the results of our other diagnostics: precipitation in MetUM-GA3 is temporally and spatially intermittent, while precipitation in GEOS5 is temporally and spatially persistent. We have added a paragraph to Section 3.1 of our revised manuscript that discusses this figure and its implications.

We thank the reviewer for this suggestion, which has helped to provide further visual evidence of our main conclusions.

p. 8, L25, "The bi-modal 1D histogram suggests that most deep convection in MetUM-GA3 is strong.": It is not clear how "strong" or stronger than what? Please add more explanations.

We agree with the reviewer that this statement is not clear. We mean that most deep convection in MetUM-GA3 is as intense as it possibly can be, at that horizontal resolution and scientific configuration of the model. We have modified this sentence to state that "most deep convection in MetUM-GA3 is as strong as possible, given the horizontal resolution and scientific configuration of the model."

p. 9, L10: "Despite having the finest horizontal resolution" should be replaced by "Because having the finest horizontal resolution"?

We disagree with the reviewer's suggestion. As our explanation in the sentence following the one that the reviewer quoted states, we would expect (naïvely, perhaps) that finer horizontal resolution would increase spatial correlations, particularly when those correlations are measured as a function of the native gridpoint. A finer-resolution model has a smaller physical distance between gridpoints, which means that a precipitation feature of the same physical dimension would span more gridpoints in a finer-resolution model than in a coarser-resolution model. Thus, one would expect the finer-resolution model to have a higher spatial correlation, measured in native gridpoints. MetUM-GA3 has the lowest spatial correlation and the finest spatial resolution, which is at odds with this expectation. Therefore, we have maintained our use of "despite".

p. 9, L13: After "The lag-1 correlation at the central gridpoint is slightly negative", add "for MetUM-GA3" for readability.

We agree with the reviewer. We have added "for MetUM-GA3" to the end of this sentence.

p. 9, L19: Delta is superscript. Please correct such that "0.5–1.5Δx".

We agree with the reviewer. We have corrected the formatting of the Δ symbol.

p. 10, L33, "While there were no observation-based constraints on timestep rainfall": This statement is incorrect. We can use the ground radar data for very high-spatial and temporal resolution of precipitation, such as 1 km and 10 min. We can also use satellite radar data for high-spatial distribution of precipitation, such as PR of TRMM or DPR of GPM. The authors should add discussions on using and analyzing such high-resolution radar data for evaluations of precipitation in future directions.

We agree with the reviewer. We were referring specifically to the lack of verifying observations for similar spatial domains and time periods to those used in our analysis of the 2-day hindcast data from the MJO inter-comparison project. We acknowledge that there are high-resolution observations, such as

those the reviewer cites, that could be compared against model simulations, assuming care was taken to perform those comparisons at similar spatial and temporal resolutions, based on the results from our study. In fact, the comparison of high-resolution, convection-permitting model simulations against radar data is an area of ongoing research.

We have modified the sentence in question to read: "While there are no observation-based constraints on timestep rainfall for similar spatial domains and temporal periods as the model data analysed here . . . ". We believe this statement is accurate.

Further, we have added a short paragraph to the Discussion section that discusses the use of high-resolution radar data to validate simulated timestep precipitation, as the reviewer suggests. This is the second paragraph of the Discussion section (Section 4) in our revised manuscript.

p. 10, L34, "Both TRMM and CMORPH produce histograms that are broader than the models? histograms and which peak at heavier precipitation rates.": These observation also have biases especially for lighter rainfalls. The authors should add remark on the biases of the observations in the earlier sections such as in the methodology.

We agree with the reviewer. We have added a sentence to Section 2 to note this bias: "Both products have been shown to under-detect light rainfall rates (e.g., Tian et al., 2010)." Further, we note that we commented on the under-detection of light rainfall on page 11, line 34 of the original manuscript. We have added the Tian et al. (2010) reference above to that sentence as well.

p. 11, L8, "(dashed line on Fig.s 4a)": "Fig.s 4a should be "Fig. 4a".

We agree with the reviewer. We have removed the erroneous "s" after "Fig.".

p. 11, L15, "Conversely, models with more persistent timestep precipitation (e.g., GEOS5, MRI-AGCM, CAM5 and MIROC5) display greater intermittency for 3-hr means.": To clarify the sentence, please add "when Fig. 9 is compared with Fig. 4". It is not clear how great the intermittency is. The authors should quantify the intermittency.

We agree with the reviewer's first point. We have added the statement "(compare Fig. 9 to Fig. 4)" to the end of the sentence in question.

We also agree with the reviewer's second point. We have added metrics for spatial coherence and temporal persistence to our revised manuscript (see response to "General Comments" above). These metrics clearly demonstrate that averaging from timestep to 3-hr means increases the intermittency in models with more-persistent timestep precipitation (see Table 3 in revised manuscript). We have added the following sentence to the end of the paragraph in question: "Table 3 confirms that temporal averaging on the native grid reduces inter-model variations in the temporal persistence summary metric, by increasing values for models with relatively low scores (e.g., MetUM-GA3, CanCM4, CNRM-AM) and reducing the values for models with relatively high scores (e.g., CAM5, MIROC5, MRI-AGCM3)."

Please note that we revised the sentence in question to read "Conversely, models with more persistent timstep precipitation (e.g., GEOS5, MRI-AGCM, CAM5 and MIROC5) display reduced persistence when data are averaged to 3-hr means" to prevent the mis-understanding that these models were the most intermittent at the 3-hr among all the models considered. In fact, these models are still the most persistent models, as our new summary metrics show, but they show reduced persistence (and lower values of our metric) for 3-hr data than for timestep data.

p. 11, L28-29, "SPCAM3, ECEarth3 and CanCM4 are perhaps closest to TRMM and CMORPH, but are still more persistent.": Again, it is not clear how these models are close to the observations. Please consider quantification of the persistency.

We agree with the reviewer. As above, we have created new metrics to quantify temporal persistence and spatial coherence, which can be found in Table 3 of the revised manuscript. These metrics demonstrate

that SPCAM3, ECEarth3, CanCM4 and CNRM-AM show temporal persistence in 3-hr precipitation that is most similar to TRMM and CMORPH, but that all models are too persistent with respect to both datasets. We have added references to our summary metrics to the paragraph in question, noting that these four models show values closest to the satellite-derived observations.

p. 11, L34, "We note that there are also differences between TRMM and CMORPH over this short period: CMORPH displays more frequent light precipitation than TRMM, which has been shown to under-detect light rainfall (Huffman et al., 2007, e.g.,). TRMM is more intermittent than CMORPH.": The authors should note why these differences come from between the two observations. "(Huffman et al., 2007, e.g.,)" should be "(e.g., Huffman et al., 2007)".

We disagree with the reviewer's first point. The purpose of our manuscript is not to understand the difference between the TRMM and CMORPH datasets. Because we do not have direct access to the algorithms used to produce the TRMM 3B42 and CMORPH products, we cannot investigate the reasons for the differences in temporal coherence between the two datasets. Any hypotheses would be mere conjecture that would not be worthy of publication. Instead, we compare these datasets only to provide a measure of observational uncertainty in our diagnostics. We have added a sentence to Section 2 to clarify this: "We employ two observation-based datasets to provide a measure of observational uncertainty in our diagnostics."

We agree with the reviewer's second point. We have corrected this errors by moving the "e.g.," to the beginning of the parenthetical citation.

p. 12, L3, "All models display higher correlations": Add "at 3h interval (Fig. 10b)" at the end of this sentence.

We agree with the reviewer. We have added "When using 3-hr data" to the start of the sentence in question and a reference to Fig. 10b at the end of the sentence.

p. 12, L11, "Spatial averaging reduces timestep intermittency in all models (Fig. 11).": The observations of TRMM and CMORPH should also be added to Fig. 11.

We disagree with the reviewer. It is not possible to show data for TRMM and CMORPH in Fig. 11, because Fig. 11 shows timestep data from models, not 3-hr data. Comparing timestep data (12–60 minutes) from models to 3-hr data from TRMM and CMORPH is not valid. Fig. 14 in our manuscript compares 3-hr data from TRMM and CMORPH to 3-hr data from models, when all datasets have also been interpolated to a 5.6°×5.6° horizontal grid. This is the equivalent of Fig. 11 for 3-hr data; it provides a fair comparison between models and satellite-derived observations.

p. 12, L18-20, "This suggests that using a common horizontal grid or a common timescale does not necessarily create a fair comparison between models, due to differences in the number of points or timesteps, respectively, that are combined to create the average.": It is not appropriate to say "a fair comparison", because it is not clear in what sense the fair comparison means. Use of a common horizontal grid or a common timescale has its importance for some purposes. Please rephrase this sentence.

We agree with the reviewer. We were referring specifically to the diagnostics we produced in this study. Of course, interpolating to a common grid or timescale is appropriate for some purposes, outside the remit of this study. We have amended this sentence to read "For the purposes of these diagnostics, using a common horizontal grid or a common timescale does not necessarily create a fair comparison between models . . . ".

p. 12, L24, "the comparison of Fig. 4a and Fig. 11a suggests that MetUM-GA3 likely has only a few precipitating gridpoints . . . ": This sentence is not clear. Please explain what the authors want to mean.

We do not know exactly which aspect of this sentence the reviewer finds unclear. Our point is that when one averages over a broad region, such as our 5.6°×5.6° boxes, two models can produce the same precipitation rate (or spectrum of rates) from different combinations of precipitation frequency and intensity on

the native gridscale. For example, Model A might produce very infrequent, but very heavy precipitation at its native resolution. Model B might produce very frequent, but light precipitation. When data from these models are averaged across a 5.6°×5.6° box, the average precipitation rates might be the same. In our study, MetUM-GA3 is like Model A. Within each 5.6°×5.6° box at any given time, there are very few precipitating gridpoints, but those gridpoints show heavy precipitation. Our new Fig. 7—which the reviewer suggested—demonstrates this well.

We have revised this sentence to read "For instance, the comparison of Fig. 4a and Fig. 12a suggests that MetUM-GA3 likely has only a few precipitating gridpoints in each 5.6°×5.6° region, but that those points show very heavy precipitation (e.g., 90–130 mm day^{-1}), as indicated in Fig. 8a."

p. 12, L39, "in Figs. 7a and 12a) implies": Delete ")".

We agree with the reviewer. We have corrected this error by deleting the extraneous right-hand parenthesis.

p. 14, L9, "Although there are no verifying observations for our timestep data": As mentioned before (p. 4, L5-6), the ground radar data can be used for verification of the timestep data.

We agree with the reviewer. Again, we were referring to the lack of verifying observations for similar spatial domains and time periods as the ones covered by the 2-day hindcast dataset we analysed. We have modified the sentence in question to read: "Although there are no verifying observations for the model timestep data that cover comparable spatial and temporal domains ...."

p. 14, L33-34: "Fig. 4a" and "Fig. 4b" should be "Fig. 7a" and "Fig. 7b", respectively.

We agree with the reviewer. We have corrected this error by replacing "Fig. 4a" and "Fig. 4b" with "Fig. 7a" and "Fig. 7b", respectively.

p. 14, Section 4 Discussion: For understanding of properties of cumulus convection schemes, single column models (SCM) have been widely used. Especially, SCM under a radiative-convective equilibrium (RCE) condition is a useful framework for understanding the timestep behaviors. For example, Satoh and Hayashi (1992, J. Atmos. Sci.), Takata and Noda (1997, J. Meteor. Soc. Japan) for SCM in RCE. Please add discussions on the above aspects of using a SCM for understanding of intermittency.

We agree with the reviewer. We have added a short paragraph to the Discussion section of our revised manuscript (second paragraph of Section 4), in which we discuss the potential use of single-column model simulations to investigate the causes of undesirable intermittency in simulated precipitation, including the references that the reviewer pointed out. However, SCM simulations can only address errors in the sub-gridscale physics, not errors in the coupling between that physics and the resolved dynamics. Hence, we have added the following sentence to the revised manuscript: "Model development efforts to reduce or remove undesirable intermittency may involve single-column model experiments, in which the effects of changes in sub-gridscale physics can be isolated from feedbacks through the resolved dynamics (e.g., Satoh and Hayashi, 1992; Takata and Noda, 1997; Woolnough et al., 2010), although we stress that physics–dynamics coupling may have a substantial effect on the model behaviors and diagnostics presented here."

ASoP (v1.0): A set of methods for analyzing scales of precipitation in general circulation models

Nicholas P. Klingaman[1], Gill M. Martin[2], and Aurel Moise[3]

[1]National Centre for Atmospheric Science–Climate and Department of Meteorology, University of Reading, Earley Gate, P.O. Box 243, Reading, Berkshire RG6 6BB, United Kingdom
[2]Met Office, Exeter, United Kingdom
[3]Bureau of Meteorology, Melbourne, Australia

Correspondence to: Nicholas P. Klingaman (nicholas.klingaman@ncas.ac.uk)

Abstract. General circulation models (GCMs) have been criticized for their failure to represent the observed scales of precipitation, particularly in the tropics where simulated daily rainfall is too light, too frequent, and too persistent. Previous assessments have focused on temporally or spatially averaged precipitation, such as daily means or regional averages. These evaluations offer little actionable information for model developers, because the interactions between the resolved dynamics and parameterized physics that produce precipitation occur at the native gridscale and timestep.

We introduce a set of diagnostics (ASoP1) to compare the spatial and temporal scales of precipitation across GCMs and observations, which can be applied to data ranging from the gridscale and timestep to regional and sub-monthly averages. ASoP1 measures the spectrum of precipitation intensity, temporal variability as a function of intensity, and spatial and temporal coherence. When applied to timestep, gridscale tropical precipitation from ten GCMs, the diagnostics reveal that far from the "dreary" persistent light rainfall implied by daily mean data, most models produce a broad range of timestep intensities that span 1–100 mm day^{-1}. Models show widely varying spatial and temporal scales of timestep precipitation. Several GCMs show concerning quasi-random behavior that may influence alter the spectrum of atmospheric waves. Averaging precipitation to a common spatial (≈600 km) or temporal (3-hr) resolution substantially reduces variability among models, demonstrating that averaging hides a wealth of information about intrinsic model behavior. When compared against satellite-derived analyses at these scales, all models produce features that are too large and too persistent.

1 Introduction

Advances in supercomputing power continue to enable refinements in the resolutions of general circulation models (GCMs) used to simulate the effects of climate variability and anthropogenic climate change. As GCMs have become better able to resolve regional-scale boundary features (e.g., orography, coastlines), the scientific community has paid increasing attention to these models' representations of local and regional hydrological extremes (e.g., Dai, 2006; Wilcox and Donner, 2007; Rosa and Collins, 2013), including the sensitivity of those extremes to climate change (e.g., Trenberth, 2011; Kharin et al., 2013; Pendergrass and Hartmann, 2014; Westra et al., 2014). Robust projections of local and regional changes in extremes with anthropogenic warming are essential to underpin decisions on adaptation strategies; accurate predictions of these extremes in

response to natural climate variability are critical for preserving lives and livelihoods, for example through emergency response and anticipatory aid efforts, particularly on sub-seasonal–seasonal scales.

Despite refinements in resolution and efforts to revise the treatment of sub-gridscale processes such as deep convection, climate models are criticized routinely for their inability to represent the observed frequency, intensity and persistence of precipitation. Dai (2006) compared daily precipitation in 18 GCMs from the Third Coupled Model Intercomparison Project (CMIP3) against satellite-derived analyses from the Tropical Rainfall Measuring Mission (TRMM) dataset across 50°S–50°N. The CMIP3 models produced precipitation too frequently, particularly light precipitation (< 10 mm day^{-1}), but did not produce heavy precipitation (> 20 mm day^{-1}) frequently enough. Models performed similarly poorly when compared against gridded gauge data over land (Sun et al., 2006). Wilcox and Donner (2007) obtained similar results at the sub-daily scale, demonstrating that 30-min averaged rainfall (sampled every 3h) from the Geophysical Fluid Dynamics Laboratory model was biased towards low intensities relative to TRMM. Revisions to the convective parameterization, particularly the closure and the triggering function, increased heavy precipitation frequency and reduced light precipitation frequency. Stephens et al. (2010) employed observations from the CloudSat spaceborne cloud-profiling radar to show that although contemporary GCMs produced reasonable seasonal and annual precipitation accumulations, these accumulations arose from highly biased daily precipitation distributions: models produced precipitation far too frequently and far too lightly. The strong preference for persistent, light daily accumulations led the authors to call the GCMs' simulated world "dreary". Such biases lead to erroneously large moisture recycling over land, with consequences for the simulation of the global hydrological cycle (e.g., Trenberth, 2011; Demory et al., 2014).

More recently, Koutroulis et al. (2015) found that GCMs from the Fifth Coupled Model Intercomparison Project (CMIP5) had improved somewhat in their daily precipitation distributions relative to their CMIP3 counterparts, particularly through an increase in the frequency of intense precipitation and a reduction in the overall frequency of precipitation. Hirota and Takayabu (2013) showed improved skill for 1–5 day precipitation extremes in CMIP5 relative to CMIP3. However, Rosa and Collins (2013) concluded that CMIP5 GCMs still produced 3-hr rain rates of 1–10 mm day^{-1} too frequently over the southeastern United States, compared to gridded gauge data. When the models did produce heavier events, those events were too persistent.

Although the studies above highlight a heightened focus on the GCM representations of hydrological extremes—which are inherently small-scale, short-lived features—most evaluation of GCM precipitation focuses on gross spatial (e.g., regional averages) and temporal (e.g., monthly and seasonal means) characteristics (e.g., Phillips and Gleckler, 2006; Bollasina and Ming, 2013; Li and Xie, 2014; Mehran et al., 2014). Where attention is paid to shorter-term variability, studies have adopted a phenomenological approach, analyzing precipitation associated with synoptic features such as mesoscale fronts and convective systems (e.g., Brown et al., 2010; Catto et al., 2013; Van Weverberg et al., 2013) or sub-seasonal modes such as the Madden–Julian oscillation (MJO; e.g, Hung et al., 2013). Yet the processes that produce precipitation in GCMs—the interactions between the sub-gridscale parameterizations and the resolved dynamics—function on the native gridscale and timestep of the models, not on a 3-hr or daily mean basis or on a regional average. Although it is often hypothesized that biases in the distributions of spatially and/or temporally averaged precipitation are the result of errors at the gridpoint, timestep level, few studies have examined the spatial and temporal characteristics of precipitation at these most fundamental scales. In isolated

single-column model experiments, convective parameterizations have been shown to produce highly intermittent timestep precipitation (e.g., Stirling and Stratton, 2012), but it is not clear how, or even if, this behavior influences the distributions of precipitation at larger and longer scales. Information about the spatial and temporal characteristics of gridscale precipitation are far more useful for informing parameterization development than information about regional biases in seasonal, or even daily, accumulations.

The dearth of studies focused on the gridscale and timestep may be due to a lack of data, since large GCM intercomparison efforts such as CMIP5 do not collect timestep output to limit the volume of data produced. However, a recent model-evaluation project focused on the MJO (Klingaman et al., 2015) collected timestep data from ten GCMs for a limited number of short hindcast simulations (Xavier et al., 2015). Xavier et al. (2015) found that models differed considerably in the degree of timestep-to-timestep precipitation variability over a $5^\circ \times 5^\circ$ region of the equatorial Indian Ocean, as computed by the root-mean-squared difference of area-averaged timestep precipitation, but did not connect this variability to other scales or examine the spectra of rainfall intensities. There was no relationship between timestep precipitation variability and MJO fidelity (Klingaman et al., 2015).

Another reason for the lack of attention to timestep precipitation may be a scarcity of suitable diagnostics to compare the characteristics of precipitation variability among models, and between models and observations, across spatial and temporal scales. Previous studies have focused mainly on frequency distributions of precipitation intensities, computed mainly at the model gridscale but often on time-averaged or selectively sampled data (e.g., one 30-min model timestep per 3-hr). While these results are useful, they do not consider the coherence of precipitation features in space and time. ~~Such diagnostics~~ Diagnostics of precipitation coherence require sampling many gridpoints and timesteps, which can be computationally cumbersome when working with high-frequency, fine-resolution data.

In this manuscript, we introduce diagnostics designed to describe precipitation variability and scale interactions in observations and models across a range of spatial and temporal scales. These diagnostics were developed with a view to condensing large data volumes from sub-daily output of O(10)km-scale GCMs into a set of measures of precipitation frequency, intensity and spatial and temporal coherence, to improve understanding of observed rainfall variability and compare simulated and observed precipitation characteristics across a range of scales. The diagnostics form a small software package entitled "Analyzing Scales of Precipitation", version 1.0 (ASoP1). In section 2, we describe the ASoP1 diagnostics, then introduce the MJO hindcast dataset mentioned above. In section 3 we demonstrate how the diagnostics can be used to discern and evaluate model behavior by applying them to the MJO hindcast dataset as well as to satellite-derived precipitation analyses. We discuss our results in section 4 and summarize our findings in section 5.

2 Diagnostics and datasets

As this manuscript focuses on novel diagnostics of precipitation, we devote section 2.1 to a thorough explanation of our methods, including examples using satellite-derived precipitation analyses. To demonstrate the ability of these diagnostics to compare precipitation characteristics among a range of model configurations, we apply the ASoP1 diagnostics to sub-daily

tropical precipitation from ten models from the "Vertical structure and physical processes of the Madden–Julian oscillation" model-evaluation project, described in section 2.2 and shown in Table 1, as well as 3-hr data from two satellite-derived analyses: TRMM 3B42 product, version 7A (Kummerow et al., 1998; Huffman et al., 2007, 2010, ; hereafter "TRMM") and the National Oceanic and Atmospheric Administration Climate Prediction Center Morphing Technique, version 1.0 (Joyce et al., 2004, ; hereafter "CMORPH"). We employ two observation-based datasets to provide a measure of observational uncertainty in our diagnostics. Both products are derived from a combination of infrared and microwave sounders~~and calibrated against gauge data.~~. The TRMM 3B42 algorithm combines available precipitation data from microwave sounders, then fills the gaps in this dataset by merging precipitation data from infrared sounders, which have been first calibrated against the microwave sounders over a longer time period. TRMM 3B42 is calibrated on a monthly basis against gauge data from the Global Precipitation Climatology Centre. The 3B42 algorithm uses many microwave sounders, not only the TRMM Microwave Imager; "TRMM" refers to the source of the data, not the instrument. CMORPH combines precipitation data from microwave sensors only, but fills the gaps between microwave satellite overpasses by advecting the precipitation field using vectors derived from infrared-based cloud observations. Both products have been shown to under-detect light rainfall rates (e.g., Huffman et al., 2007; Tian et al., 2010). We use TRMM and CMORPH in the domain 60°–160°E, 10°S–10°N for two periods in boreal winter 2009–10, the choice of which is described in section 2.2.

TRMM and CMORPH have a native horizontal resolution of 0.25°×0.25°, which is finer than any of the models analyzed. Because the diagnosed spatial and temporal scales of precipitation will vary with horizontal resolution, we use an area-weighted averaging method to interpolate TRMM and CMORPH to a 1.25°×1.25 grid, which is approximately the median resolution of the models (147 km). A robust validation of any one model would require averaging TRMM and CMORPH to the model's native resolution, or preferably to a common resolution coarser than the model's native grid, as our results suggest. However, model validation is not the purpose of our study, so for clarity of presentation we compare the models to the 1.25° TRMM and CMORPH data to indicate observed scales of precipitation at a resolution comparable to, but not exactly equal to, the models' resolution. The example diagnostics below demonstrate the effects of horizontal resolution on the scales of precipitation, using 0.25° and 1.25° CMORPH data.

We discuss the diagnostics first, as they are designed to be applied to any model or observed dataset at scales ranging from the model timestep to a sub-seasonal average, and from the gridscale to O(1000 km) regions, depending on the phenomena and scales of interest. The results we show in section 3 for timestep and 3-hr precipitation are only one example use of these diagnostics. In all our diagnostics, we scale precipitation rates to mm day^{-1}, since this units is commonly used in other studies. However, it should be remembered that a fixed value in mm day^{-1} equates to various rainfall intensities depending on the temporal scale considered (e.g., a 20-min timestep or a 3-hr average).

2.1 Methods

2.1.1 Precipitation spectra and contributions to total precipitation

To examine the precipitation intensity distribution on a given temporal or spatial scale, and its sensitivity to temporal and spatial averaging, we compute the contributions of discrete bins of precipitation intensity to the total precipitation at a gridpoint. These contributions can be expressed as either a precipitation rate, where the sum across all bins gives the total precipitation rate, or as a fraction of the total precipitation rate, where the sum across all bins is unity. In the latter case, the result is a spectrum that shows the relative importance of precipitation events in a given intensity bin to the total precipitation, while the former also includes contributions from the frequency of each precipitation rate. We use 100 bins (b; mm day^{-1}), for which the edges are defined by:

$$b_i = e^{\left\{\ln(0.005)+\left[i\cdot\frac{(\ln(120)-\ln(0.005))^2}{59}\right]^{\frac{1}{2}}\right\}} \quad (1)$$

where i is the number of the bin and $\ln(x)$ is the natural logarithm of x. We add a further lower bin edge at 0.0 to ensure that a histogram computed using these bins sums to the number of valid data points in the sample.

The calculations can be performed for any input grid and temporal resolution. By calculating these contributions at each gridpoint in a region, we produce maps of the contributions of precipitation intensity bins to the total precipitation at each gridpoint. Examples of these for 3-hr TRMM and CMORPH 1.25° data are shown in Fig. 1. These contributions can then be accumulated over a sub-region and plotted as one-dimensional (1D) histograms, allowing easy comparison of the spectral characteristics of rainfall for the sub-region across temporal or spatial scales and between datasets.

2.1.2 Two-dimensional histograms

To diagnose the behavior of satellite-derived and simulated precipitation between consecutive temporal intervals at a fixed gridpoint, we construct two-dimensional (2D) histograms of gridpoint precipitation in temporal interval t against precipitation at the same gridpoint in the next interval $t + \Delta t$, where Δt is the sampling frequency of the input data. Gridpoint precipitation is binned, using bins that give a roughly uniform distribution for 2000–2012 TRMM analyses over an extended tropical Warm Pool domain (10°S–10°N, 60°–160°E), while also maintaining a pseudo-logarithmic scale. The 2D histograms are normalized by the total number of data points, such that the integral of the normalized histograms is unity. Figs. 2a,b show examples of this diagnostic for CMORPH 0.25° and 1.25° data. For a given cell (i,j), the value shown is the joint probability of precipitation at a gridpoint in intensity bin i during temporal interval t and precipitation at the same gridpoint in intensity bin j during temporal interval $t + \Delta$t. Averaging from 0.25° to 1.25° resolution slightly reduces the frequency of very heavy precipitation (> 180 mm day^{-1}) and near-zero precipitation, while slightly increasing the frequency of rates in between. Averaging also increases the probability of persistent precipitation in consecutive 3-hr intervals, as there are higher probabilities towards the central diagonal and lower probabilities along the axes in Fig. 2b relative to Fig. 2a.

2.1.3 Correlations with distance and lag

Correlations of precipitation in space and time indicate the typical scales of convective features. To compute these, we divide the analysis domain into non-overlapping sub-regions of 7×7 gridpoints. We ~~find~~ select the central point in each region and extract the timeseries of precipitation. Computing the instantaneous correlation between the precipitation timeseries at each point in the sub-region and the central point, then averaging the resulting 7×7 correlation maps across all sub-regions in the analysis domain, creates a field of composite lag-0 correlations like those shown in Figs. 2~~b~~ c and 2~~c~~ d for CMORPH data. As expected, the correlations decrease with distance away from the central point. Correlations decrease more quickly along the diagonal axes, for which distances are greater, than along the major axes; correlations also decrease more quickly in the meridional direction than in the zonal direction, likely because the prevailing winds in our extended tropical Warm Pool domain are zonal. Correlations are lower for the 1.25° than for 0.25° CMORPH data, which is expected as each 1.25° gridpoint represents a 5x greater physical distance than at 0.25°.

While the composite correlation maps are useful, we are interested in both spatial and temporal scales of precipitation, which requires computing lagged correlations. It would be cumbersome to produce a set of composite correlation maps, one for each lag, for each datasets in this study. Instead, we developed a summary diagram that combines information about the spatial and temporal correlations of precipitation, based on the same 7×7 sub-regions. The construction of this diagram is described below; examples using CMORPH are shown in Figs. 2e and 2f.

We compute the distance (d; in km) between each point in the sub-region and the central point and convert this distance into units of Δx (the longitudinal grid spacing at the equator). We bin the gridpoints in the sub-region by their distance from the central point in Δx units, using bins of width Δx starting from 0.5Δx (e.g., 0.5Δx $< d \leq$ 1.5Δx, 1.5Δx $< d \leq$ 2.5Δx). ~~For completeness, we include a~~ We omit the bin of 0 $<$ d $\leq$ 0.5Δx, ~~although~~ as no datasets in our study have a grid with Δy $\leq$ 0.5Δx in the tropics (where Δy is the latitudinal grid spacing). We treat the central gridpoint as a separate bin.

Within each distance bin, we compute the average correlation at a range of lags between the precipitation timeseries of gridpoints in that bin and the central gridpoint in the sub-region. For each 7×7 sub-region, these computations result in a matrix of correlations with distance and time, as shown in ~~Fig~~Figs. 2~~e~~e and 2f. Note that all correlations are computed against the central point at lag=0. We average these matrices across all 7×7 sub-regions. For the central point (marked "Centre" in Figs. 2e and 2f), the result is simply the average of the autocorrelations of the central points in all sub-regions. At lag-0, the result is similar to the average of the correlations shown in ~~Fig~~Figs. 2~~b~~ c and 2d within each distance range. ~~For the ranges shown here, the CMORPH 0.25° correlations decline more quickly with time than with space.~~ At all gridpoints, the precipitation timeseries is no longer statistically significantly correlated with itself (at p=0.05) after six hours (lag=2). At lag=2, the correlations at all distances are essentially uniform, including at the central point, which suggests that all spatial information from the lag=0 precipitation field has been lost (i.e., if the gridpoints in the lag=2 field were randomly swapped, one could not identify which was the central gridpoint). The CMORPH 1.25° data demonstrates that averaging increases correlations with time, due to the greater physical distance represented by each gridpoint (Fig. 2f), as significant correlations are maintained until nine hours (lag=3).

2.1.4 Comparisons among models and between models and analyses

The example of CMORPH 0.25° and 1.25° data demonstrates that the correlations in Figs. 2c–f are difficult to compare across datasets with different resolutions, because they are expressed as functions of the native gridscale. To compare spatial scales of precipitation features across resolutions, we repeat the method described in section 2.1.3 but using sub-regions defined by physical distance, rather than a number of gridpoints. To ensure that we include correlations to a distance of at least 2.5Δx in the coarsest-resolution models considered here (Table 1), and to optimize the number of sub-regions relative to the size of the domain, we divide the analysis domain into sub-regions of approximately 1500$\times$1500 km, rounded to a distance equal to a whole number of gridpoints in the input dataset. Thus, the size of the sub-regions varies slightly from one dataset to the next. Within each sub-region, we bin the gridpoints by distance from the central gridpoint in units of Δx and compute correlations as in section 2.1.3. A higher-resolution model or dataset will contain more gridpoints in each sub-region, and so have more distance bins, than a lower-resolution model or dataset, but this method allows a cleaner comparison between datasets of different resolutions.

For each dataset, we compute the minimum, median and maximum physical distance from the central point within each distance bin. This allows us to construct a graph of the correlation with physical distance at lag=0. Fig. 3a shows an example comparing TRMM and CMORPH at 0.25° and 1.25° resolutions. Each point represents one distance bin, plotted at the median distance for that bin; the horizontal solid lines span the minimum and maximum distance for that bin. Spatial averaging slightly increases correlations at the same distance for both TRMM and CMORPH. Estimates of the spatial scale of precipitation features from a finer-resolution dataset will be lower than those from a coarser-resolution version of the dataset.

To compare temporal correlations of precipitation, we use the mean auto-correlation of precipitation at all gridpoints within the analysis domain. Fig. 3b shows an example of this analysis, again for TRMM and CMORPH; each point represents one timestep in the input dataset. Spatial averaging ~~also~~ increases estimates of the temporal scale of precipitation features.

We also create summary metrics of temporal and spatial coherence in precipitation. First, we compute quartiles of precipitation at each gridpoint, using only rates $>$ 1 mm day^{-1} to prevent near-zero precipitation values from dominating the lowest quartile. Computing quartiles separately at each gridpoint accounts for spatial variations in the distributions of precipitation across the analysis domain. To measure temporal coherence, we compute the probabilities that, at the same gridpoint and on consecutive timesteps, upper-quartile precipitation (U) is followed by upper-quartile precipitation [in probability notation, p(U|U)]; lower-quartile precipitation (L) is followed by lower-quartile precipitation [p(L|L)]; upper-quartile precipitation is followed by lower-quartile precipitation [p(L|U)]; and lower-quartile precipitation is followed by upper-quartile precipitation [p(U|L)]. When computing p(L|L) and p(L|U), the lowest quartile is expanded to include rates $\leq$ 1 mm day^{-1} for only the second timestep, to account for transitions to near-zero precipitation. In other words, p(L|U) is the probability that upper-quartile precipitation is followed by precipitation below the threshold for the lowest quartile, including rates $\leq$1 mm day^{-1}. High values of p(U|U) and p(L|L) demonstrate temporal persistence; high values of p(L|U) and p(U|L) demonstrate temporal intermittency. As a metric of coherence (M), we combine these probabilities using

$$M = 0.5 \times [p(U|U) + p(L|L) - p(L|U) - p(U|L)] \quad (2)$$

High values of M represent greater persistence. The factor of 0.5 ensures that the range of possible M values spans -1.0 to 1.0. Negative M indicates that intermittency is more common than persistence; positive values indicate that persistence is more common than intermittency.

To measure spatial coherence, we divide the analysis domain into non-overlapping regions of 3×3 gridpoints, in the same manner as for the 7×7 regions in section 2.1.3. In each region, we select the central gridpoint and find all instances of upper-quartile precipitation. For those timesteps, we compute the probability of upper-quartile precipitation at the eight other gridpoints in the 3×3 region [p(U|U)], as well as the probability of lower-quartile precipitation [p(L|U)]. We then compute similar probabilities for timesteps with lower-quartile precipitation at the central gridpoint [p(L|L) and p(U|L)]. As for the temporal persistence metric, we expand the lowest quartile to include values < 1 mm day^{-1} when assessing precipitation at neighboring gridpoints. Finally, we compute M using (2) above. As for temporal persistence, high values of M represent greater spatial coherence, while low values represent spatial intermittency.

Table 3 shows that for TRMM and CMORPH, averaging from 0.25° to 1.25° horizontal resolution reduces the spatial coherence of precipitation, but increases the temporal persistence of precipitation. The reduction in spatial coherence is due to the metric being computed on the native gridscale of the input dataset, rather than physical distance; as above, the results in Fig. 3a show that the 1.25° datasets have larger spatial features than the 0.25° datasets. It is not practical to use physical distance in our summary spatial coherence metric, as selecting a fixed physical distance would be problematic for certain applications and regions (e.g., close to steep topography, land/sea contrasts).

2.1.5 Spatial and temporal averaging

To assess the sensitivity of sub-daily precipitation variability to the choice of spatial and temporal scale, we compute many of the above diagnostics using not only precipitation at a model's native gridscale and timestep, but also precipitation that has been averaged in time or space or both. For all models, we average timestep precipitation to 3-hr means for ease of comparison with TRMM and CMORPH. For all models and TRMM and CMORPH, we use an area-weighted method to average gridscale precipitation onto a common 5.6°×5.6° grid that is approximately four times coarser than the coarsest-resolution models used in this study. Using this grid, rather than the native grid of the coarsest-resolution models, ensures that all models are subject to some degree of spatial averaging, which our results show can substantially impact sub-daily precipitation statistics.

2.2 Models

We obtained gridpoint, timestep precipitation data from ten of the 12 models that participated in the two-day hindcast component of the "Vertical structure and physical processes of the Madden–Julian oscillation (MJO)" model-evaluation project (Xavier et al., 2015). The project was organised by the Global Atmospheric Systems Studies (GASS) panel, the Years of Tropical Convection (YoTC) and the MJO Task Force. We did not obtain data from the European Centre for Medium-range Weather Forecasts (ECMWF) Integrated Forecasting System, because ECMWF submitted hourly averages rather than timestep data. We omitted the Pacific Northwest National Laboratory configuration of the Weather Research and Forecasting model, because an incomplete dataset was archived. Table 1 lists the models, their timesteps and native horizontal resolutions, as well as ref-

erences with further details on their formulations. In the model-evaluation project, each model performed 48-hour hindcasts, initialized once per day from 00Z ECMWF operational analyses during two strong MJO events in boreal winter 2009–10. There are 22 start dates per event: 20 October–10 November 2009 and 20 December 2009–10 January 2010. To reduce the effects of model adjustment from the ECMWF analyses, we removed the first 12 hours of each hindcast, as in Xavier et al. (2015), to leave 1584 hours of data (36 hours$\times$44 hindcast dates) for each model. Data are available for all gridpoints in 10°S–10°N and 60°–160°E. Each of the two hindcast periods contains an active MJO that propagates from the Indian Ocean to the West Pacific, such that most gridpoints in the domain experience a transition from active to suppressed or suppressed to active MJO conditions during each event. This reduces the likelihood that our results depend upon MJO phase. TRMM and CMORPH are analysed at 1.25° resolution for the same period.

While the period of the hindcast experiments is relatively short, this is the only known multi-model dataset of timestep output from full-physics GCMs on the models' native grids. In addition, the dataset includes tendencies of temperature, humidity and winds from the individual sub-gridscale physical parameterizations in these models. While we do not consider these tendencies here, they represent a useful avenue for further research into the causes of the model behavior shown here.

For the GASS/YoTC models, TRMM and CMORPH, Table 2 gives the number of 7$\times$7 sub-regions, the number of 1500$\times$1500 km sub-regions and the dimensions of the 1500$\times$1500 km sub-regions in native gridpoints.

3 Results

In all figures, we order the GASS/YoTC models alphabetically by abbreviation (Table 1) except that we place MetUM-GA3 first. MetUM-GA3 often displays behaviour distinct from the other models. Because of the attention paid to MetUM-GA3 in our discussion, and because MetUM is the subject of our future work, we choose to separate this model to emphasize its unique behaviour.

3.1 Behavior on the native grid and timestep

Two-dimensional histograms (section 2.1.2) reveal that the GASS/YoTC models vary considerably in their levels of temporal variability in gridpoint, timestep tropical precipitation (Fig. 4). On these diagrams, high probabilities along the central diagonal indicate persistent precipitation rates on consecutive timesteps at the same gridpoint. Low probabilities along the diagonal and high probabilities in the lower-right and upper-left quadrants, close to the axes, identify intermittent precipitation at a gridpoint: high probabilities in the lower-right quadrant indicate that moderate or heavy precipitation is often followed by light or no precipitation, while high probabilities in the upper-left indicate that light or no precipitation is often followed by moderate or heavy precipitation. MetUM-GA3 ~~is~~ has by far the most ~~"temporally intermittent" model by~~ intermittent precipitation this measure. The 1D histogram suggests that MetUM-GA3 oscillates between lighter ($<$ 9 mm day^{-1}) and heavier ($>$ 30 mm day^{-1}) rain rates, with almost no instances of moderate rates (9–30 mm day^{-1}). Heavier precipitation almost never persists for more than one timestep, while light or near-zero precipitation is much more likely to be followed by light or near-zero precipitation on the next timestep. This behavior suggests that when MetUM-GA3 triggers convection, if that

convection is strong, the convection alters the thermodynamic profile such that it is highly unlikely that strong convection will be triggered on the next timestep. The bi-modal 1D histogram suggests that most deep convection in MetUM-GA3 is ~~strong~~as strong as possible, given the horizontal resolution and scientific configuration of the model.

Among the other models, CNRM-AM, GISS-E2, SPCAM3, ECEarth3 and CanCM4 show some degree of timestep intermittency in precipitation. Unlike MetUM-GA3, however, all of these models have higher values on the central diagonal than away from it (i.e., the most likely value of precipitation at one gridpoint and timestep is the value of precipitation at the same gridpoint on the previous timestep). CNRM-AM and CanCM4 show behavior most similar to MetUM-GA3, with probabilities on the abscissa and ordinate axes that are nearly as high as those on the central diagonal. In CanCM4, the 2D PDF is almost uniform for rates < 60 mm day^{-1}, suggesting random behavior; rates ≥ 60 mm day^{-1} are more persistent.

In contrast, GEOS5, MRI-AGCM, CAM5 and MIROC5 are ~~"temporally persistent" models~~ models with more persistent precipitation, in which gridpoint precipitation at one timestep is highly correlated with precipitation at the next timestep. These models maintain this behavior across the spectrum of intensity, such that even very heavy precipitation is much more likely to be followed by very heavy precipitation than by light or near-zero precipitation. This implies that in these models, strong convection does not result in a stable profile that inhibits convection on the next timestep. We note that there is no correspondence between the length of the model timestep and temporal intermittency in precipitation: of the six models with 30-min timesteps (Table 1), three ~~are relatively intermittent~~ produce relatively intermittent precipitation (CNRM-AM, GISS-E2 and SPCAM3), while three ~~are relatively persistent~~ produce relatively persistent precipitation (MRI-AGCM, CAM5 and MIROC5).

To evaluate spatial coherence of timestep precipitation and temporal variability at lags > 1 timestep, we use the diagnostic of the average correlation with distance and lag described in section 2.1.3 (Fig. 5). All models show decreasing correlations with distance from the central point and with temporal lag, as expected. Despite having the finest horizontal resolution, MetUM-GA3 produces the lowest correlations of any model in space and time. All other aspects being equal, horizontal resolution should increase spatial correlations when measured as a function of Δx, as seen in Fig. 2 for CMORPH. The lag-1 correlation at the central gridpoint is slightly negative for MetUM-GA3. The correlation then increases at subsequent lags, reaching a maximum at lag-4. CNRM-AM also shows a lag-1 minimum in the auto-correlation of timestep precipitation, but the lag-1 correlation is still strongly positive in that model (0.683). MetUM-GA3 also shows very low spatial coherence: the instantaneous correlation of precipitation at the central gridpoint with the precipitation at points 0.5–1.5Δx away is not statistically significant at the 10% level (r=0.13; p$\sim$0.15). This implies that timestep precipitation in MetUM-GA3 cannot be reliably predicted from precipitation at neighboring gridpoints at the same timestep, or from previous timesteps at the same gridpoint; it is quasi-random. CanCM4 displays similar behavior, with a instantaneous correlation of only 0.17 between the central point and points 0.5–1.5~~Δ~~Δx away. CanCM4 has a Δx that is approximately five times longer than MetUM-GA3, however. Indeed, with the exception of MetUM-GA3, models with coarser horizontal resolution (GISS-E2, SPCAM3, CanCM4) tend to show lower spatial correlations than models with finer resolution (GEOS5, CAM5, MRI-AGCM). This may be expected, since the physical area of the 7×7 boxes considered for this diagnostic will be far larger in the coarser-resolution models than in the finer-resolution ones. Naïvely, one would expect a larger area to have more spatially heteorogeneous large-scale forcing, and hence less coherent precipitation.

This hypothesis is difficult to confirm with such a wide variety of GCMs—which differ in many respects beyond horizontal resolution (e.g., sub-gridscale parameterizations)—and suggests the need for resolution-based sensitivity experiments with a single model.

Fig. 6 compares the spatial and temporal scales of timestep precipitation in the GASS/YoTC models. On the native grid and timestep, MetUM-GA3 is clearly an outlier, with by far the lowest spatial (Fig. 6a) and temporal (Fig. 6b) coherence in precipitation. Only MetUM-GA3 and CNRM-AM show a lag-1 minimum in the auto-correlation of timestep precipitation; in MetUM-GA3 the correlation remains lower the other models for the remainder of the 3-hr period considered.

With the exception of MetUM-GA3, the models exhibit similar rates of decline in precipitation coherence with increasing distance. Models which show relatively higher correlations in first distance bin (CAM5, GEOS5, MIROC5, MRI-AGCM3) tend to have relatively higher correlations at longer distances; models with relatively lower correlations (CanCM4, CNRM-AM, SPCAM3) also maintain that behavior. The same is true for the decrease in correlation with increasing lag. There is a clear link between spatial and temporal coherence in these models: models which show relatively higher spatial coherence also tend to show relatively higher temporal coherence, and vice versa.

The summary metrics of coherence confirm these results (Table 3). The models identified as having the most temporally persistent precipitation based on 2D histograms and auto-correlations—CAM5, GEOS5, MRI-AGCM3 and MIROC5—also emerge as the models with the highest values of the temporal persistence summary metric. These models also show relatively high spatial coherence, as does ECEarth3. MetUM-GA3, CanCM4 and CNRM-AM display the least temporal persistence of precipitation, as well as low spatial coherence of precipitation; SPCAM3 also shows low spatial coherence of precipitation, but relatively higher temporal persistence of precipitation. MetUM-GA3 is the only model to produce a negative value for the spatial coherence metric, suggesting that spatial intermittency in precipitation is more common than spatial coherence.

To provide sample visualizations of the spatial and temporal character of precipitation in models with intermittent and persistent precipitation, we show example timeseries and maps of timestep, gridpoint precipitation from MetUM-GA3 and GEOS5 (Fig. 7). We selected MetUM-GA3 as our analysis suggests it is by far the model with the most intermittent precipitation; we selected GEOS5 because its timestep and horizontal resolution are similar to MetUM-GA3 (Table 1), but it produces more persistent precipitation in time and space (Table 3). Timeseries of precipitation at a gridpoint in the middle of the Indian Ocean (0°, 90°E), for a forecast initialised during the active phase of the first MJO event (4 November 2010), confirms that MetUM-GA3 produces temporally intermittent precipitation (Fig. 7a). All timesteps with precipitation rates > 5 mm day^{-1} also have rates > 100 mm day^{-1}. By contrast, GEOS5 produces some precipitation on nearly all timesteps, with only a few timesteps exceeding 100 mm day^{-1} (Fig. 7b). Maps of instantaneous precipitation rates in the two models show that MetUM-GA3 precipitation is also intermittent in space, and dominated by gridpoints with precipitation rates > 100 mm day^{-1} (Fig. 7c), while GEOS5 precipitation is far more continuous with a broader distribution of intensities (Fig. 7d).

Even after removing MetUM-GA3 as an outlier, it is obvious that the remaining models exhibit a broad range of spatial and temporal coherence in their precipitation features on the native grid and timestep. Next, we consider whether these timestep and gridpoint characteristics influence the models' behavior at on longer and larger scales.

3.2 Effects of temporal averaging

We begin by considering the impact of averaging from timestep to 3-hr data on the distributions of precipitation intensity in the GASS/YoTC models, using histograms of the fractional contribution from each of the precipitation bins defined in eq. 1 to the total precipitation (Fig. 8). As in Fig. 4, the timestep histograms demonstrate the range of precipitation intensities produced by these models, with MetUM-GA3 generating almost all of its precipitation from intense timestep events $>$100 mm day^{-1} (Fig. 8a). Maps of contributions to the average precipitation rate confirm that this is true across most of the domain (Fig. 9a), not just in the regionally-aggregated statistics. Most of the other models produce the majority of their precipitation from 10–100 mm day^{-1} timestep events, including ECEarth3, which favors the 10–50 mm day^{-1} intensity range over most of the Warm Pool (Fig. 9b). There are no relationships between the preferred intensity of precipitation and timestep length or horizontal resolution.

When all data are averaged to a common 3-hr resolution, the differences between the models reduce considerably (Fig. 8b). While averaging barely affects the histogram for some models (CAM5, CNRM-AM, GEOS5, MIROC5), for other models averaging shifts the PDF considerably (CanCM4, MetUM-GA3, SPCAM3). For this latter set of models, the dominant effect is to reduce the contributions from heavy precipitation ($>$100 mm day^{-1}) and increase the contributions from moderate precipitation (10–50 mm day^{-1}). This is the expected result for averaging a random process, but it is not clear that timestep precipitation within a 3-hr window should be random. The effect is clearly greatest for MetUM-GA3, which has very low temporal coherence of timestep precipitation and a short timestep (i.e., more timesteps are averaged together to produce the 3-hr average). The models least affected by temporal averaging are those with persistent timestep precipitation rates (e.g., CAM5, MIROC5). All models produce a narrower histogram with a sharper peak for 3-hr means than for timestep data. Combined with the reduction in the inter-model spread with temporal averaging, the narrower histograms demonstrates that analysing only averaged precipitation hides a wide variety of model behavior at the timestep level.

For a ~~temporally intermittent model~~ model with temporally intermittent precipitation like MetUM-GA3, temporal averaging can have a powerful effect on conclusions about the dominant precipitation rate. MetUM-GA3 produces nearly all of its precipitation in timesteps with $\geq$ 100 mm day^{-1} rates, but the right column of Fig. 9a demonstrates that if one analysed only 3-hr data, one would believe that tropical precipitation fell almost exclusively in 10–50 mm day^{-1} events. This could have important implications for parameterization development. This issue does not affect a ~~temporally persistent model~~ model with more persistent precipitation like ECEarth3, which at the timestep and 3-hr scale generates most of its precipitation from 10–50 mm day^{-1} events (Fig. 9b).

While there ~~were~~ are no observation-based constraints on timestep rainfall for similar spatial domains and temporal periods as the model data analysed here, at the 3-hr scale we compare gridpoint data from the models to 1.25° TRMM and CMORPH data. Both TRMM and CMORPH produce histograms that are broader than the models' histograms and which peak at heavier precipitation rates. This suggests that, over the relatively short hindcast period, all of the models produce their precipitation from too-frequent, too-light 3-hr events (Fig. 8b). However, as noted above, the model 3-hr histograms do not represent the full range of timestep precipitation rates.

Two-dimensional histograms of 3-hr data (Fig. 10) demonstrate that averaging reduces, but does not eliminate, the variations in temporal intermittency among the models seen in the timestep data. Models with higher temporal intermittency in timestep precipitation (e.g., MetUM-GA3, CNRM-AM, SPCAM, CanCM4) show reduced intermittency for 3-hr means, with higher values along the central diagonal and lower values along the axes. The reduced intermittency is particularly striking for MetUM-GA3, in which the bi-modal PDF of timestep precipitation (dashed line on Fig.~~s~~ 4a) becomes considerably more uniform. This implies that the frequent moderate 3-hr values (9–30 mm day^{-1}) arise from a linear combination of timesteps of near-zero and very heavy ($>$ 30 mm day^{-1}) precipitation, since these moderate precipitation values are completely missing from the timestep PDF (Fig. 4a). This supports the results from the 1D histograms (Fig. 8). The reduced intermittency at the 3-hr scale may be most clear in MetUM-GA3 because the timestep intermittency was so strong, or because of the shorter timestep in MetUM-GA3 relative to the other ~~intermittent models~~ models with intermittent precipitation, which increases the effect of the averaging because more timesteps are combined.

Conversely, models with more persistent timestep precipitation (e.g., GEOS5, MRI-AGCM, CAM5 and MIROC5) display ~~greater intermittency for~~ increased intermittency when data are averaged to 3-hr means ~~.~~ (compare Fig. 10 to Fig. 4). As for the ~~more-intermittent models~~ models with more intermittent precipitation, this can be explained with a "regression to the mean" argument: averaging several timesteps of a ~~less-intermittent model~~ model with less intermittent precipitation introduces variability into the 3-hr timeseries from the occasional deviation of the timestep precipitation away from the central diagonal in Fig. 4. These models show much smaller changes in the 1D histogram between the timestep and 3-hr scales, relative to ~~the intermittent models~~ models with more intermittent precipitation, which suggests that the 3-hr values arise from many timesteps with rates close to the 3-hr mean. Again, this supports the results from the 1D histograms. Table 3 confirms that temporal averaging on the native grid reduces inter-model variations in the temporal persistence summary metric, by increasing values for models with relatively low scores (e.g., MetUM-GA3, CanCM4, CNRM-AM) and reducing the values for models with relatively high scores (e.g., CAM5, MIROC5, MRI-AGCM3).

With the exception of MetUM-GA3, it is clear that models with longer timesteps tend to show greater intermittency in 3-hr precipitation. This is likely because in these models, fewer timesteps have been combined to create the 3-hr mean. Since sub-daily precipitation data (e.g., 3-hr means or timestep values sampled every 3-hr) are often used in studies of extreme events, such as tropical cyclones, it is worth noting this apparent correlation between model timestep length and variability in precipitation rates, which could introduce sampling uncertainty into these studies. We find no relationship between spatial resolution and temporal intermittency in 3-hr precipitation.

~~All models show much~~ The 2D PDFs suggest that all models show greater persistence in 3-hr precipitation than TRMM (Fig. 10a) and CMORPH (Fig. 10b), which is confirmed by the temporal persistence metric (Table 3). SPCAM3, CNRM-AM, ECEarth3 and CanCM4 are perhaps closest to TRMM and CMORPH, but ~~are still more persistent~~still produce more persistent precipitation than the satellite-derived analyses. The variations in spatial resolution among the models, and between the models and TRMM and CMORPH, make it difficult to compare the 2D PDFs and the summary metrics directly, however. Section 3.4 revisits the comparison between the models and the satellite-based analyses using precipitation data that has been interpolated to a common horizontal grid. We note that there are also differences between TRMM and CMORPH over this short period:

CMORPH displays more frequent light precipitation than TRMM, ~~which has been shown to~~ although previous studies have shown that both products under-detect light rainfall ~~(Huffman et al., 2007, e.g.,); TRMM is more intermittent than CMORPH~~ (e.g., Huffman et al., 2007; Tian et al., 2010); TRMM precipitation is less persistent than CMORPH (Table 3). Even given the uncertainty in the satellite-based analyses, however, all models show greater temporal persistence than the analyses.

Fig. 11 summarizes the impact of temporal averaging on the spatial scale of precipitation features. Averaging increases the spatial scale for all models, but most dramatically for MetUM-GA3, although that model still has relatively low spatial correlations. ~~All~~ When using 3-hr data, all models display higher correlations (greater coherence) than TRMM and CMORPH at distances shorter than 300 km, after which the TRMM and CMORPH correlations become statistically insignificant at the 5% level ($r < \sim 0.2$; Fig. 11b). CMORPH has slightly larger precipitation features than TRMM, as well as higher values of the spatial coherence summary metric (Table 3).

~~We do not show lagged auto-correlations for~~ For model data, lagged correlations of 3-hr precipitation ~~to avoid complications from the strong~~ (i.e., as in Fig. 6b but for 3-hr data) over a 36-hr window were dominated by the overly strong and regular diurnal cycle of ~~tropical precipitation , which is often poorly represented in~~ precipitation in the models, which manifested itself in our diagnostics as a pronounced peak in the correlations at a 24-hr lag (not shown). TRMM and CMORPH displayed a much weaker and broader peak across lags of 18–30 hr, suggesting greater day-to-day variability in the timing of the diurnal maximum in tropical maximum in the satellite-observations than in the models.

3.3 Effect of spatial averaging

To investigate the effects of spatial averaging, we area-average timestep data from all models to a common 5.6°×5.6° (approximately 620 km) horizontal grid that is four times the resolution of the coarsest models (SPCAM and CanCM4). Spatial averaging ~~reduces timestep intermittency~~ increases timestep persistence of precipitation in all models~~(~~, as shown by the 2D PDFs in Fig. 12 ~~)~~and the temporal persistence summary metrics in Table 3. As for averaging to 3-hr means, spatial averaging reduces ~~the~~ intermittency most strongly in those models which either (a) have high levels of intermittency at the gridscale (e.g., MetUM-GA3, CNRM-AM, SPCAM3) or (b) have finer native resolution, as more gridpoints are averaged to create each 5.6°×5.6° box (e.g., MetUM-GA3, GEOS5, CAM5, MIROC5). Both (a) and (b) apply to MetUM-GA3, so it is not surprising that spatial averaging substantially reduces temporal intermittency. At the 5.6° scale, MetUM-GA3 is still one of models with the most intermittent ~~models~~precipitation, but while it was an outlier at the gridpoint scale, it is now largely indistinguishable from the other ~~intermittent models~~ models with intermittent precipitation (e.g., CanCM4, GISS-E2, SPCAM3). The other ~~intermittent models~~ models with intermittent precipitation have a much coarser native resolution than MetUM-GA3, however (Table 1), which means that those models have not "benefited" from combining as many gridpoints. ~~This suggests that~~ For the purposes of these diagnostics, using a common horizontal grid or a common timescale does not necessarily create a fair comparison between models, due to differences in the number of points or timesteps, respectively, that are combined to create the average.

At the 5.6° scale, the 1D histograms of precipitation (dashed lines on Fig. 12) and the spectra of precipitation contributions (Fig. 13a) become strikingly similar among the models, despite the variety of timestep lengths (12–60 min). This suggests

that, when averaged over a broad enough region, these models produce similar spectra of timestep precipitation, even though the spectra of native-gridpoint precipitation varies considerably. For instance, the comparison of Fig. 4a and Fig. 12a suggests that ~~MetUM-GA3 likely has only a few precipitating gridpoints in~~ within each 5.6°×5.6° region, MetUM-GA3 has only a few gridpoints with non-zero precipitation, but that those points show very heavy precipitation (e.g., 90–130 mm day^{-1}), as indicated in Fig. 8a. In MetUM-GA3, the difference between a 5.6°×5.6° region with relatively light (e.g., 5 mm day^{-1}) and relatively heavy (e.g., 30 mm day^{-1}) precipitation is likely that the latter region has a few more gridpoints with very heavy precipitation than the former. By contrast, the comparison of Fig. 4f and Fig. 12f, and the similarity of the MIROC5 spectra in Figs. 8a and 13a ~~)~~ implies that MIROC5 has many precipitating gridpoints in each 5.6°×5.6° region, most of which have a precipitation rate similar to the average for the region. Models for which spatial averaging results in little change in the 2D and 1D histograms are likely to have more spatially coherent precipitation, at least within a ∼5° region, than models for which spatial averaging causes large changes in the character of timestep precipitation.

Fig. 14 compares the impact of spatial averaging on the temporal scales of precipitation features across models. As for temporal averaging, spatial averaging increases the temporal scale of precipitation in all models, but most notably in ~~intermittent models~~ models with intermittent precipitation such as MetUM-GA3 and CNRM-AM. Still, there is substantial inter-model spread in the auto-correlations. Some models (CAM5, GEOS5, MIROC5, MRI-AGCM3) show nearly perfect correlations, while others (CanCM4, CNRM-AM, MetUM-GA3, SPCAM3) show relatively smaller values. Even when averaging fairly large (≈360,000 km^2) regions, the lag-1 minima in MetUM-GA3 and CNRM-AM remain, showing the strong effects of timestep intermittency from self-limiting convection in those models.

We do not show correlations with distance for the spatially averaged data, as those correlations rely on 1500 km×1500 km sub-regions that contain only ≈4 5.6°×5.6° gridpoints. Larger sub-regions are not possible as the dataset spans only 20° latitude. However, this could be done for larger (e.g. global) datasets or for individual models with higher spatial resolution. Instead, we show our summary metrics for spatial coherence, which use regions of only 3×3 gridpoints (Table 3). Most models show similar coherence for timestep, 5.6°×5.6° resolution data; as for timestep, gridpoint precipitation, MIROC5 and CAM5 are the most coherent, with MetUM-GA3 and CanCM4 the least coherent. Note that it is not appropriate to compare the spatial coherence metrics for the native-grid and 5.6°×5.6° data, since the metric is sensitive to the resolution of the input data.

3.4 Effect of spatial and temporal averaging

Combining spatial and temporal averaging produces the cleanest comparison possible among the models and between the models and TRMM and CMORPH, but at the expense of masking the timestep and gridpoint variability from Fig. 4. Histograms of precipitation intensity show that 3-hr averaging of the spatially-averaged data further reduces the differences between the models' intensity spectra, as well as between the models and TRMM and CMORPH (Fig. 13). Most models still produce too-frequent precipitation at lighter rates than TRMM and CMORPH, even when analyzed on a common grid (Fig. 13c), a result which is emphasized by taking the difference between the models' spectra and the CMORPH spectrum (Fig. 13d). Differences between the models and TRMM are similar (not shown). All models except GISS-E2 generate too much of their precipitation from light events and too little from heavy events.

At the 5.6° and 3-hr scale, the models also produce remarkably similar levels of temporal ~~coherence~~persistence (Table 3), as well as highly similar precipitation PDFs (Fig. 15). All models show low levels of intermittency, with maxima in the 2D histogram along the central diagonal and minima along the ordinate and abscissa. The similarities are particularly notable given the wide variety of behavior seen at the timestep and gridpoint level. Even MetUM-GA3 produces a 2D histogram a precipitation PDF that agrees well with the other models. At these scales, the models also agree with TRMM and CMORPH, although ~~all models remain~~ CAM5, GEOS5, MIROC5 and MRI-AGCM3 have slightly more persistent precipitation than the satellite-based analyses ~~.~~ (Table 3). Most models show higher values of the spatial coherence metric than TRMM and CMORPH, suggesting that precipitation features are still too broad.

The convergence of model behavior at the ∼600 km, 3-hr scale, combined with the close agreement with TRMM and CMORPH, implies a natural compensation in these models at the gridpoint and timestep level between the spatial and temporal intermittency in precipitation and the precipitation PDF. In other words, it seems that the models "adjust" the frequency and intensity of precipitation at their native resolutions to maintain an appropriate distribution of tropical precipitation at the broader ∼600 km and 3-hr scales. We hypothesize that these broader scales represent those at which ~~these models maintain radiative–convective equilibrium in the tropics, in which the average convective heating balances the average radiative cooling~~simulated convection is in balance with the synoptic-scale, dynamical systems that produce precipitation, predictions of which should be highly similar among the models in the short, 2-day hindcasts we analysed. At finer and shorter scales, the models have sufficient degrees of freedom to produce the broad spectrum of behavior seen in Fig. 4 and Fig. 5~~, while maintaining this equilibrium at longer and larger scales~~. Therefore, it appears that the nature of the timestep, gridpoint variability does not substantially affect the distribution of precipitation or its variability at the ∼600 km and 3-hr scales. However, it remains unclear whether a model's timestep, gridpoint behavior influences other aspects of the simulation (e.g., through interactions between convective heating and the resolved dynamics). We discuss this further in section 4.

4 Discussion

Our diagnostics reveal that analyzing temporally or spatially averaged precipitation can hide a wealth of information about model behavior on the native gridscale and timestep. This is true even for relatively small averaging scales, such as 3-hr means or 2×2 gridboxes (our 5.6°× 5.6° regions were 4× the gridscale of the coarsest resolution models in our dataset). Analysis of gridpoint, timestep precipitation is critical for developing sub-gridscale parameterizations, since these are the scales at which the parameterizations interact with the resolved dynamics. Such analysis can identify potentially undesirable characteristics, such as the strong spatial and temporal intermittency in convection in MetUM-GA3. Nearly all of the convection in MetUM-GA3 is very strong, producing precipitation rates >100 mm day^{-1} on a timestep (Fig. 8a); also, convection is often isolated to a single gridpoint and timestep (Fig. 5a). Although there are no verifying observations for ~~our timestep data~~ the model timestep data that cover comparable spatial and temporal domains, it is difficult to believe that this behavior is representative of oceanic tropical convection. These intense, isolated precipitation features must be associated with intense, isolated column heating. Over a sequence of timesteps, this behavior produces a "checkerboard"-style spatial pattern of heating that shifts from one

timestep to the next as gridpoint convection triggers quasi-randomly. It is not clear whether the model dynamics respond to this strong gridscale heating, or only to the average heating over several gridpoints and timesteps, but gravity waves triggered by the intermittent heating in one column may influence the likelihood of convection at neighboring gridpoints on subsequent timesteps, disrupting convective organization and the propagation of waves with longer periods and larger horizontal scales (e.g., Kelvin waves, the MJO). Understanding the controls on spatial and temporal intermittency in MetUM convection, as well as the influences of that intermittency on the model dynamics, tropical convective variability and the mean state, are all active areas of ~~further~~ research inspired by our diagnostics.

Future research should also seek to validate timestep model convection against high-resolution observations, through comparison of model data against ground-based or space-borne precipitation radar measurements (e.g., from the Global Precipitation Measurement mission). These comparisons must take care to analyse observed and simulated data at comparable spatial and temporal scales, given our results on the effects of spatial and temporal averaging on the distributions and coherence of precipitation. Model development efforts to reduce or remove undesirable intermittency may involve single-column model experiments, in which the effects of changes in sub-gridscale physics can be isolated from feedbacks through the resolved dynamics (e.g., Satoh and Hayashi, 1992; Takata and Noda, 1997; Woolnough et al., 2010), although we stress that physics–dynamics coupling may have a substantial effect on the model behaviors and diagnostics presented here.

Although MetUM-GA3 is the ~~most intermittent model~~ model with the most intermittent precipitation in our study, CNRM-AM, CanCM4, GISS-E2, ECEarth3 and SPCAM3 display varying degrees of intermittency (Fig. 4). It is likely that all of those models have a self-limiting character to their convective parameterizations, such that the effect of convection on one timestep reduces the probability of convective for one or several subsequent timesteps. Preliminary analysis of MetUM-GA3 (not shown) suggests that convection on one timestep produces downdraft cooling that stabilizes the vertical temperature profile near the lifting condensation level (LCL), the stability across which is used in the diagnosis of deep convection (i.e., to diagnose deep convection, the parcel must be able to ascend through the LCL). Although instability may remain aloft, the model is unable to convect on subsequent timesteps until the profile again becomes unstable at the LCL. There are a variety of mechanisms by which a parameterization can be self-limiting, which will depend on the precise design of the parameterization; a detailed examination of the convective parameterizations of ten GCMs is outside the scope of this study, but our analysis of this behavior may be of interest to individual modeling centers to understand and improve their parameterizations.

On the gridpoint and timestep scale, the worlds simulated by these models are definitely not "dreary" (e.g., Stephens et al., 2010), at least over the Warm Pool domain considered here. In most models, the total precipitation consists of a variety of timestep rates that span 1–100 mm day^{-1}, with most precipitation falling in timesteps with precipitation rates > 10 mm day^{-1} (Fig. 48a). Only when the timestep data are averaged to 3-hr means do the precipitation spectra begin to collapse to be lighter (Fig. 48b) and more persistent (Fig. 10) than in the satellite-derived analyses. The narrower spectra arise from the tendency for one timestep with heavy precipitation to be followed by several timesteps with no precipitation; the persistence of 3-hr rain rates suggests that the timestep intermittency occurs consistently in each 3-hr window. These results imply that the self-limiting character of a model's convection, displayed through temporal intermittency in timestep precipitation, prevents the model from producing enough consecutive timesteps of heavy precipitation, or enough consecutive timesteps of no precipitation, to generate

a broader distribution of 3-hr mean rates. An observer stationed on an island in the Warm Pool in many of these models would be constantly dodging intense, short-lived downpours, not standing in the persistent light rain implied by past studies' analysis of 3-hr or daily mean data.

Much of our analysis has focused on timestep and gridpoint data from GCMs, the formulations of which include spatial and temporal smoothing (either implicitly or explicitly), as well as truncation errors, both of which lead to an underestimation of energy on the smallest resolved scales. Previous studies have found that the "effective resolution" of a GCM—the scales at which the truncation and smoothing have no effect, or zero power—is several times the native resolution (e.g., Skamarock, 2004; Frehlich and Sharman, 2008; Larsén et al., 2012), such that the timestep, gridpoint data are unreliable and should be discarded. While we do not argue with the conclusions of those studies, we believe that it remains important to examine the characteristics of native-resolution data for several reasons: (a) to inform parameterization development, as discussed above; (b) to understand the effects of intermittency on these scales, however under-resolved, because that intermittency may influence the larger and longer scales in a GCM; and (c) because despite previous conclusions on effective resolution, the scientific community is increasingly using gridscale, instantaneous output from models with ever-finer horizontal resolution to study extreme events and their responses to natural variability and anthropogenic climate change (e.g., Kendon et al., 2014).

We used 2-day hindcasts from the "Vertical structure and physical processes of the MJO" model-evaluation project, which is the only known source of timestep, gridpoint precipitation data from many contemporary models. However, this dataset has limitations. First, only two sets of 22 2-day forecasts were performed, each for a case study of an MJO event in boreal winter 2009–10. Although each set of forecasts samples the MJO active and suppressed phase, limiting the possibility of sensitivity to MJO phase, there is an active MJO in the analysis domain throughout the dataset, which may bias the simulated precipitation characteristics. We plan to address this issue in a future study by computing our diagnostics for across an entire season of MetUM timestep data. Secondly, the spatial domain of the data is limited to the deep tropical Warm Pool; the dataset may not represent the full spectrum of tropical convection in the models or satellite-derived analyses. Thirdly, all forecasts were initialized from ECMWF analyses. Xavier et al. (2015) found this led to an initialization shock, the strength of which varied among the models. To reduce the effect of the shock, we removed the first 12 hr of each forecast, as in Xavier et al. (2015), but it is possible that our findings are influenced by the shock and may not represent the model's intrinsic behavior. Removing the first 24 hr of each forecast made only a very small difference to our results and did not affect our conclusions.

The analysis in section 3 is only one potential use of these diagnostics. Understanding the spatial and temporal characteristics of precipitation is important for a variety of applications. Computing precipitation spectra (Fig. 8) and 2D histograms (Fig. 4) for daily-mean or pentad-mean precipitation from models and observations could give insight into the simulated levels of synoptic and intraseasonal variance in a particular region, for instance the active and break periods of the major monsoons. Spatial maps of contributions from sections of the precipitation spectra (Fig. 9) could aid understanding of whether biases in simulated mean precipitation are due primarily to biases in frequency or in intensity. Spatial and temporal coherence diagnostics (Fig. 3) may provide information on convective aggregation, which is important for tropical cyclones and the MJO. All of these diagnostics could be used to compare precipitation characteristics from simulations of the same model at various horizontal resolutions, or with perturbations to one or several parameters, to assist model development and assessment. We believe that

these diagnostics will be useful primarily on sub-monthly and sub-2000 km scales, as larger and longer scales are likely dominated by the seasonal cycle rather than the individual synoptic or mesoscale systems that produce precipitation.

When comparing datasets with different spatial and temporal resolutions, it is commonplace to average all data to the resolution of the coarsest dataset. However, our results show that any spatial or temporal averaging can alter precipitation characteristics, such that for the purposes of these diagnostics it is unfair to compare a lower-resolution dataset at its native resolution to a higher-resolution dataset that has been averaged to the lower resolution. Instead, we recommend comparing the datasets at their native resolutions—to understand the behavior of each dataset—as well as at a common resolution at least 2× (in each direction) that of the coarsest dataset in space and time. This is still not a clean comparison because the effects of averaging increase with the number of points combined (up to some asymptotic limit), but at least it allows both datasets to "experience" some averaging in space and time.

5 Conclusions

We have developed a range of diagnostics to identify the spatial and temporal characteristics of precipitation in observations and GCMs; these diagnostics form a small software package, "Analyzing Scales of Precipitation" version 1.0 (ASOP1). The ASoP1 diagnostics are designed be applied to sub-monthly data at horizontal resolutions O(1000 km) or finer, to assess precipitation variability associated with phenomena ranging from individual cloud systems to mesoscale weather systems and synoptic fronts. The diagnostics are motivated by the increasing attention paid to the simulation of local and regional hydrological extremes in fine-resolution GCMs—which often requires gridscale, instantaneous precipitation data—while model evaluation has remained focused primarily on monthly and seasonal accumulations. Sub-gridscale parameterization development requires information about the spatial and temporal variability of precipitation at the native gridscale and timestep, since these are the scales at which the parameterizations operate. The ASoP1 diagnostics include 1D histograms and spatial maps of the contributions of intensity ranges to the total precipitation (e.g., Fig. 8 and Fig. 9); 2D histograms of precipitation rates at the same gridpoint on consecutive time intervals (e.g., Fig. 2a), which show the temporal persistence of precipitation; the average correlation of precipitation at a range of distances and temporal lags, correlated against precipitation at a central gridpoint (Fig. 2c), computed by dividing the analysis domain into a series of non-overlapping sub-regions (e.g., Fig. 2b); ~~and~~ average correlations as a function of either physical distance (in km) or time, with which one can compare datasets with different spatial and temporal resolutions (e.g., Fig. 3); and summary metrics that can be used to track easily the effects of changes to model resolution, physics or dynamics on the spatial and temporal coherence of precipitation (Table 3).

To demonstrate the value of these diagnostics, we apply them to ten models from the "Vertical structure and physical processes of the MJO" model-evaluation project (Table 1), which collected timestep data at the native model horizontal resolution over an extended Warm Pool domain (10°S–10°N, 60°–160°E) from 44 2-day hindcasts during two strong MJO events in boreal winter 2009–10. At the timestep and gridscale, some models produce precipitation features that are highly coherent in space and time, while others produce intermittent precipitation that resembles uncorrelated noise (Fig. 4). MetUM-GA3 is the ~~most intermittent model~~model with the most intermittent precipitation, with a weakly negative lag-1 auto-correlation of

timestep precipitation and no statistically significant correlations between precipitation at neighboring gridpoints (Fig. 6). We found no relationship between the level of intermittency and either horizontal resolution or the length of the model timestep. ~~Intermittent models~~ Models with intermittent precipitation tend to produce more of their total precipitation from very heavy events—often exceeding 100 mm day^{-1} in the case of MetUM-GA3—while models with persistent timestep precipitation, such as ECEarth3, generate more frequent precipitation with moderate intensities of 10–50 mm day^{-1} (Figs. 8 and 9). Strong and highly intermittent convection, such as that in MetUM-GA3, will be associated with strong and intermittent column heating, which may interact with the resolved dynamics, affecting the spectrum of tropical wave activity and even the mean state. The effects of this intermittency remain unclear, but are an active area of research. The fact that five of the ten GCMs in this study produce heavy timestep precipitation rates, interspersed by timesteps of little or no precipitation, contradicts the common criticism that GCMs simulate a "dreary state" in the tropics of continual light precipitation, which arose from studies that analyzed 3-hr or daily averaged precipitation (e.g., Stephens et al., 2010). In fact, many models continually produce short-lived, intense downpours throughout the Warm Pool.

Averaging timestep, gridscale data in either time (to 3-hr means) or space (to 5.6°×5.6°) considerably reduces inter-model variations in the spatial and temporal scales of precipitation (Figs. 11 and 14, Table 3), as well as in the spectra of precipitation intensities (Fig. 8) and the temporal persistence of precipitation rates (Figs. 10 and 12). This is because spatial or temporal averaging has a greater effect on intermittent precipitation than on persistent precipitation. When compared to TRMM and CMORPH satellite-derived precipitation analyses over the same period and domain, all models produce 3-hr precipitation features that are too broad and too persistent, despite the fact that many of those same models produce timestep precipitation features that are isolated in both space and time (Fig. 11). This emphasizes that averaging in either space or time can hide a wealth of information about the intrinsic behavior of GCMs.

Averaging 3-hr data from the models, TRMM and CMORPH to a common 5.6°×5.6° grid improves the agreement among the models, as well as between the models and the satellite-derived analyses (Figs. 13 and 15, Table 3). We hypothesize that the strong agreement among the models indicates that these are the scales at which the ~~models maintain radiative–convective equilibrium over the tropical Warm Pool~~convection in these models is in balance with the synoptic-scale, dynamical systems that produce precipitation. This convergence of model behavior may be enhanced by the fact that these data are from short (2-day) forecasts initialized from the same ECMWF analyses, which means ~~the models should have more similar radiative-cooling profiles than they would~~ that the representation of these dynamical systems are much more similar among models than if the data came from free-running climate simulations.

These results represent only one possible use of the ASoP1 diagnostics, which we believe will be useful for model development and evaluation at longer (e.g., daily, synoptic) and larger (e.g., regional averages) scales, as well as at the native gridpoint and timestep. In particular, these diagnostics would be ideal for understanding the effects of horizontal resolution and changes to physical parameters on the simulated spatial and temporal scales of precipitation, and for comparing the characteristics of precipitation and their representation in models in different tropical regions.

6 Code availability and requirements

The ASoP1 diagnostics package is coded in Python 2. The code is available ~~for non-commercial research use upon request from Nicholas Klingaman (nicholas.klingaman@ncas. ac. uk) or Gill Martin (gill.martin@metoffice.gov.uk).~~ under the terms of the Apache 2.0 license from the lead author's GitHub repository at https://github.com/nick-klingaman/ASoP. There are two software packages: ASoP-Spectral, which computes the 1D histograms and maps of the contributions of specific intensity bins to the total precipitation; and ASoP-Coherence, which computes the 2D PDFs, the correlations with distance and time and the spatial and temporal coherence summary metrics. The user must install several Python packages prior to running the code; a list of these is given at the top of each python code file in the package. These packages also have software dependencies. The hardware requirements for running the code will vary based on the size of the dataset the user wishes to analyze, particularly for the amount of system memory (RAM) required. The analysis shown in this manuscript was performed on a four-core desktop workstation with 32GB RAM.

7 Data availability

Data from the "Vertical structure and physical processes of the Madden–Julian oscillation" project can be obtained from the Earth System Grid Federation: https://www.earthsystemcog.org/projects/gass-yotc-mip.

TRMM 3B42 version 7A data can be obtained from http://disc.sci.gsfc.nasa.gov/TRMM.

CMORPH version 1.0 data can be obtained from ftp://ftp.cpc.ncep.noaa.gov/precip/global_CMORPH/3-hourly_025deg.

Author contributions. N. Klingaman developed the diagnostics using 2D histograms, correlations versus distance and lag and correlations in space and time. G. Martin and A. Moise developed the diagnostics using 1D histograms and maps of the contributions of intensity bins. N. Klingaman wrote the manuscript with input from all co-authors.

Acknowledgements. N. Klingaman was supported by an Independent Research Fellowship from the UK Natural Environment Research Council (NE/L010976/1). G. Martin was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101).

References

Bollasina, M. A. and Ming, Y.: The general circulation model precipitation bias over the southwestern equatorial Indian Ocean and its implications for simulating the South Asian monsoon, Clim. Dynam., 40, 823–838, 2013.

Brown, J. R., Jakob, C., and Haynes, J. M.: An evaluation of rainfall frequency and intensity over the Australian region in a global climate model, J. Climate, 23, 6504–6525, 2010.

Catto, J. L., Jakob, C., and Nicholls, N.: A global evaluation of fronts and precipitation in the ACCESS model, Aust. Meteorol. Oceanogr. Soc. J., 63, 191–203, 2013.

Dai, A.: Precipitation characteristics in eighteen coupled climate models, J. Climate, 19, 4606–4630, 2006.

Demory, M.-E., Vidale, P. L., Roberts, M. J., Berrisford, P., Strachan, J., Schiemann, R., and Mizielinski, M. S.: The role of horizontal resolution in simulating drivers of the global hydrological cycle, Clim. Dynam., 42, 2201–2225, 2014.

Frehlich, R. and Sharman, R.: The use of structure functions and spectra from numerical model output to determine effective model resolution, Mon. Wea. Rev., 136, 1537–1553, 2008.

Hazeleger, W., Wang, X., Severijns, C., Stefanescu, S., Bintanja, R., Sterl, A., Wyser, K., Semmler, T., Yang, S., van den Hurk, B., van Noije, B., van der Linden, E., and van der Wiel, K.: EC-Earth v2.2: description and validation of a new seamless earth system prediction model, Clim. Dynam., 39, 2611–2629, 2012.

Hirota, N. and Takayabu, Y. N.: Reproducibility of precipitation distribution over the tropical oceans in CMIP5 multi-model models compared to CMIP3, Clim. Dynam., 41, 2909–2920, 2013.

Huffman, G. J., Adler, R. F., Bolvin, D. T., Gu, G., Nelkin, E. J., Bowman, K. P., Hong, Y., Stocker, E. F., and Wolff, D. B.: The TRMM multi-satellite precipitation analysis: quasi-global, multi-year, combined-sensor precipitation estimates at fine scale, J. Hydrometeorol., 8, 38–55, 2007.

Huffman, G. J., Adler, R. F., Bolvin, D. T., and Nelkin, E. J.: The TRMM multi-satellite precipitation analysis (TMPA), in: Satellite rainfall applications for surface hydrology, edited by Hossain, F. and Gebremichael, M., pp. 3–22, Springer Verlag, 2010.

Hung, M.-P., Lin, J.-L., Wang, W., Kim, D., Shinoda, T., and Weaver, S. J.: MJO and convectively coupled equatorial waves simulated by CMIP5 climate models, J. Climate, 26, 6185–6214, 2013.

Joyce, R. J., Janowiak, J. E., Arkin, P. A., and Xie, P.: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution, J. Hydrometeorol., 5, 487–503, 2004.

Kendon, E. J., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C., and Senior, C. A.: Heavier summer downpours with climate change revealed by weather forecast resolution model, Nat. Clim. Chang., 4, 570–576, 2014.

Khairoutdinov, M., DeMott, C., and Randall, D.: Evaluation of the simulated interannual and subseasonal variability in an AMIP-style simulation using the CSU multiscale modeling framework, J. Climate, 23, 413–431, 2008.

Kharin, V. V., Zwiers, F. W., Zhang., X., and Wehner, M.: Changes in temperature and precipitation extremes in the CMIP5 ensemble, Climatic Change, 119, 345–357, 2013.

Klingaman, N. P., Jiang, X., Xavier, P. K., Petch, J., Waliser, D., and Woolnough, S. J.: Vertical structure and physical processes of the Madden–Julian oscillation: synthesis and summary, J. Geophys. Res. Atmos., 120, 4671–4689, 2015.

Koutroulis, A. G., Grillakis, M. G., Tsanis, I. K., and Papadimitriou, L.: Evaluation of precipitation and temperature simulation performance of the CMIP3 and CMIP5 historical experiments, Clim. Dynam., in press, doi:10.1007/s00 382–015–2938–x, 2015.

Kummerow, C., Barnes, W., Kozu, T., Shiue, J., and Simpson, J.: The Tropical Rainfall Measuring Mission (TRMM) sensor package, J. Atmos. Oceanic Technol., 15, 809–817, 1998.

Larsén, X. G., Ott, S., Badger, J., Hahmann, A. N., and Mann, J.: Recipes for correcting the impact of effective mesoscale resolution on the estimation of extreme winds, J. Appl. Meteorol. Climatol., 51, 521–532, 2012.

Li, G. and Xie, S.-P.: Tropical biases in CMIP5 multimodel ensemble: the excessive equatorial Pacific cold tongue and double ITCZ problems, J. Climate, 27, 1765–1780, 2014.

Mehran, A., AghaKouchak, A., and Phillips, T. J.: Evaluation of CMIP5 continental precipitation simulations relative to satellite-based gauge-adjusted observations, J. Geophys. Res. Atmos., 119, 1695–1707, 2014.

Merryfield, W. J., Lee, W.-S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., Ajayamohan, R. S., Fyfe, J. C., Tang, Y., and Polavarapu, S.: The Canadian seasonal to interannual prediction system. Part I: Models and initialization, Mon. Wea. Rev., 141, 2910–2945, 2013.

Neale, R. B. et al.: Description of the NCAR Atmospheric Model: CAM5.0, Tech. Rep. NCAR/TN-486+STR, National Center for Atmospheric Research, Boulder, Colorado, USA, 2012.

Pendergrass, A. G. and Hartmann, D. L.: Changes in the distribution of rain frequency and intensity in response to global warming, J. Climate, 27, 8372–8383, 2014.

Phillips, T. J. and Gleckler, P. J.: Evaluation of continental precipitation in 20th century climate simulations: the utility of multimodel statistics, Water Resour. Res., 42, W03 202, 2006.

Rienecker, M. M., Suarez, M. J., Todling, R., Bacmeister, J., Takacs, L., Liu, H.-C., Gu, W., Sienkiewicz, M., Koster, R. D., Gelaro, R., Stajner, I., and Nielsen, J. E.: The GEOS-5 data assimilation system: Documentation of version 5.0.1, 5.1.0 and 5.2.0, Tech. rep., Technical Report series on Global Modeling and Data Assimilation, 2008.

Rosa, D. and Collins, W. D.: A case study of subdaily simulated and observed continental convective precipitation: CMIP5 and multiscale global climate models comparison, Geophys. Res. Lett., 40, 5999–6003, 2013.

Satoh, M. and Hayashi, Y.-Y.: Simple cumulous models in one-dimensional radiative convective equilibrium problems, J. Atmos. Sci., 49, 1202–1220, 1992.

Schmidt, G., Kelley, M., Nazarenko, L., Ruedy, R., Russell, G. L., Aleinov, I., Bauer, M., Bauer, S. E., Bhat, M. K., Bleck, R., Canuto, V., Chen, Y.-H., Cheng, Y., Clune, T. L., Del Genio, A., de Fainchtein, R., Faluvegi, G., Hansen, J. E., Healy, R. J., Kiang, N., Koch, D., Lacis, A. A., LeGrande, A. N., Lerner, J., Lo, K. K., Matthews, E. E., Menon, S., Miller, R. L., Oinas, V., Oloso, A. O., Perlwitz, J. P., Puma, M. J., Putman, W. M., Rind, D., Romanou, A., Sato, M., Shindell, D. T., Sun, S., Syed, R. A., Tausnev, N., Tsigaridis, K., Unger, N., Voulgarakis, A., Yao, M.-S., and Zhang, J.: Configuration and assessment of the GISS ModelE2 contributions to the CMIP5 archive, J. Adv. Model. Earth Syst., 6, 141–184, 2014.

Skamarock, W. C.: Evaluating mesoscale NWP models using kinetic energy spectra, Mon. Wea. Rev., 132, 3019–3032, 2004.

Stephens, G., L'Ecuyer, T., Forbes, R., Gettlemen, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., Gabriel, P., and Haynes, J.: Dreary state of precipitation in global models, J. Geophys. Res., 115, D24 211, 2010.

Stirling, A. J. and Stratton, R. A.: Entrainment processes in the diurnal cycle of deep convection over land, Q. J. R. Meteorol. Soc., 138, 1135–1149, 2012.

Sun, Y., Solomon, S., Dai, A., and Portmann, R. W.: How often does it rain?, J. Climate, 19, 916–934, 2006.

Takata, K. and Noda, A.: The effect of cumulus convection on CO_2 induced climate change in the tropics, J. Meteorol. Soc. Japan, 75, 677–686, 1997.

Tian, Y., Peters-Lidard, C. D., and Eylander, J. B.: Real-time bias reduction for satellite-based precipitation estimates, J. Hydrometeorol., 11, 1275–1285, 2010.

Trenberth, K. E.: Changes in precipitation with climate change, Clim. Res., 47, 123–138, 2011.

Van Weverberg, K., Vogelmann, A. M., Lin, W., Luke, E. P., Cialella, A., Minnis, P., Khaiyer, M., Boer, E. R., and Jenson, M. P.: The role of cloud microphysics parameterization in the simulation of mesoscale convective system clouds and precipitation in the tropical western Pacific, J. Atmos. Sci., 70, 1104–1128, 2013.

Voldoire, A., Sanchez-Gomez, E., Salas y M??lia, D., Decharme, B., Cassou, C., Sénési, S., Valcke, S., Beau, I., Alias, A., Chevallier, M., Dq́ué, M., Deshayes, J., Douville, H., Fernandez, E., Madec, G., Maisonnave, E., Moine, M.-P., Planton, S., Saint-Martin, D., Szopa, S., Tyteca, S., Alkama, R., Belamari, S., Braun, A., Coquart, L., and Chauvin, F.: The CNRM-CM5.1 global climate model: description and basic evaluation, Clim. Dynam., 40, 2091–2121, 2013.

Walters, D. N., Best, M. J., Bushell, A. C., Copsey, D., Edwards, J. M., Falloon, P. D., Harris, C. M., Lock, A. P., Manners, J. C., Morcrette, C. J., Roberts, M. J., Stratton, R. A., Webster, S., Wilkinson, J. M., Willett, M. R., Boutle, I. A., Earnshaw, P. D., Hill, P. G., MacLachlan, C., Martin, G. M., Moufouma-Okia, W., Palmer, M. D., Petch, J. C., Rooney, G. G., Scaife, A. A., and Williams, K. D.: The Met Office Unified Model Global Atmosphere 3.0/3.1 and JULES Global Land 3.0/3.1 configurations, Geosci. Model Dev., 4, 919–941, 2011.

Watanabe, M., Suzuki, T., Oíshi, R., Komuro, Y., Watanabe, S., Emori, S., Takemura, T., Chikira, M., Ogura, T., Sekiguchi, M., Takata, K., Yamazaki, D., Yokohata, T., Nozawa, T., Hasumi, H., Tatebe, H., and Kimoto, M.: Improved climate simulation by MIROC5: Mean states, variability and climate sensitivity, J. Climate, 23, 6312–6335, 2010.

Westra, S., Fowler, H. J., Evans, J. P., Alexander, L. V., Berg, P., Johnson, F., Kendon, E. J., Lenderink, G., and Roberts, N. M.: Future changes to the intensity and frequency of short-duration extreme rainfall, Rev. Geophys., 52, 522–555, 2014.

Wilcox, E. M. and Donner, L. J.: The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model, J. Climate, 20, 53–69, 2007.

Woolnough, S. J., Blossey, P. N., Xu, K.-M., Bechtold, P., Chaboureau, J.-P., Hosomi, T., Iacobellis, S. F., Luo, Y., Petch, J. C., Wong, R. Y., and Xie, S.: Modelling convective processes during the suppressed phase of a Madden–Julian oscillation: Comparing single-column models with cloud-resolving models, Q. J. R. Meteorol. Soc., 136, 333–353, 2010.

Xavier, P. K., Petch, J. C., Klingaman, N. P., Woolnough, S. J., Jiang, X., Waliser, D. E., Caian, M., Hagos, S. M., Hannay, C., Kim, D., Cole, J., Miyakawa, T., Prithard, M., Roehrig, R., Shindo, E., Vitart, F., and Wang, H.: Vertical structure and physical processes of the Madden–Julian oscillation: Biases and uncertainties at short range, J. Geophys. Res., 120, 4749–4763, 2015.

Yukimoto, S., Adachi, Y., Hosaka, M., Sakami, T., Yoshimura, H., Hirabara, M., Tanaka, T. Y., Shindo, E., Tsujino, H., Deushi, M., Mizuta, R., Yabu, S., Obata, A., Nakano, H., Ose, T., and Kitoh, A.: A new global model model of Meteorological Research Institute: MRI-CGCM3–model description and basic performance, J. Meteorol. Soc. Japan, 90A, 23–64, 2012.

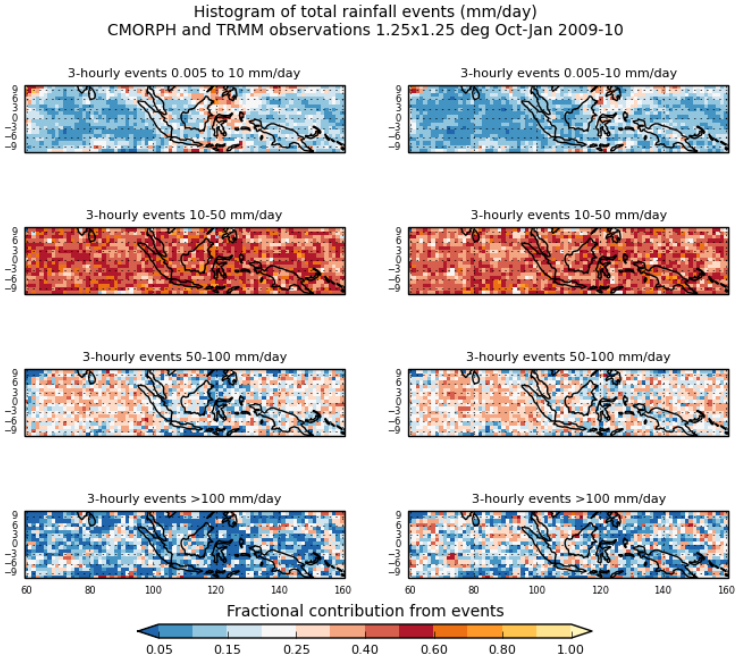


Figure 1. For (left) CMORPH and (right) TRMM 3B42 1.25° data, the fractional contribution to the total precipitation rate from ranges of intensity bins shown in the labels above each panel. For each dataset, the sum of each column is unity.

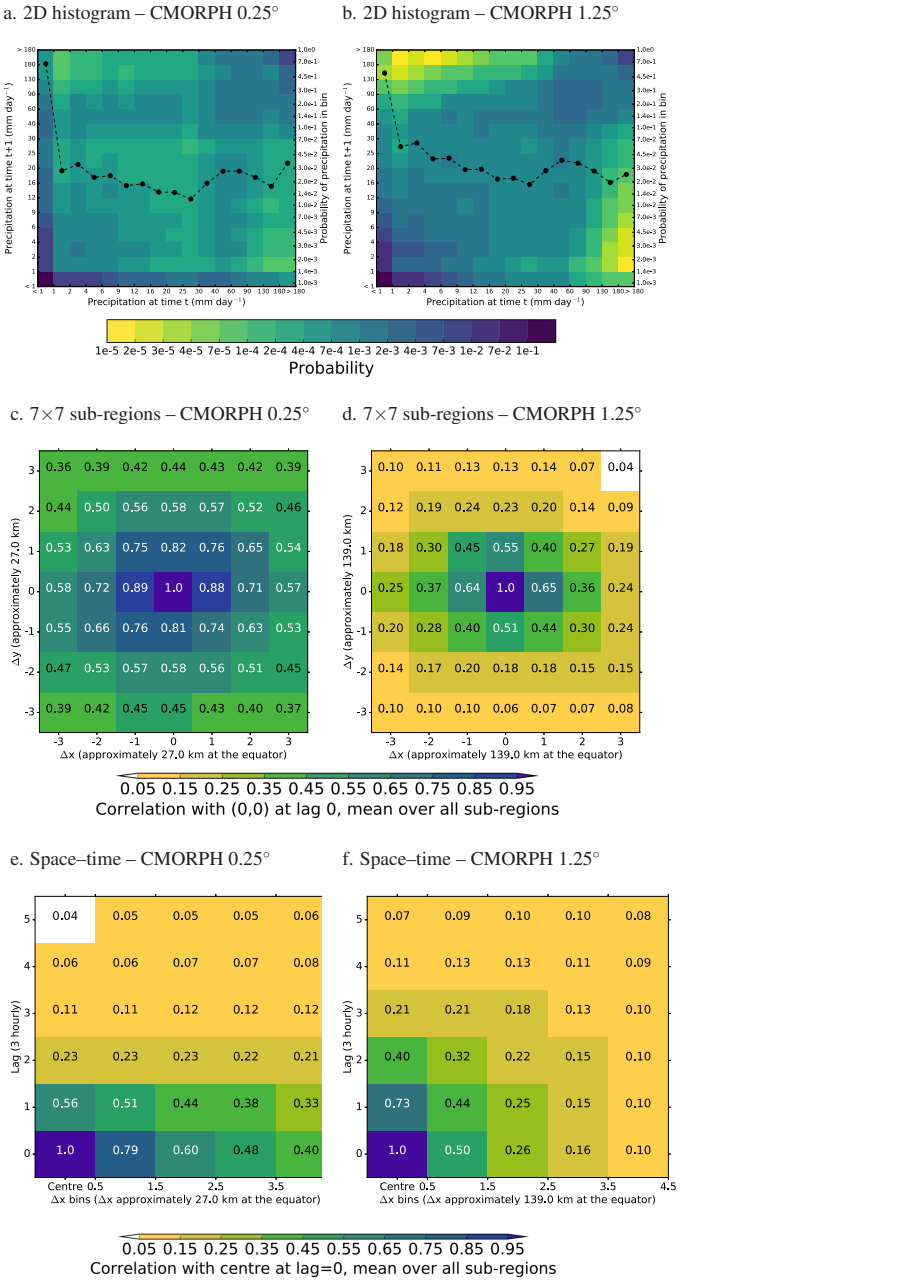
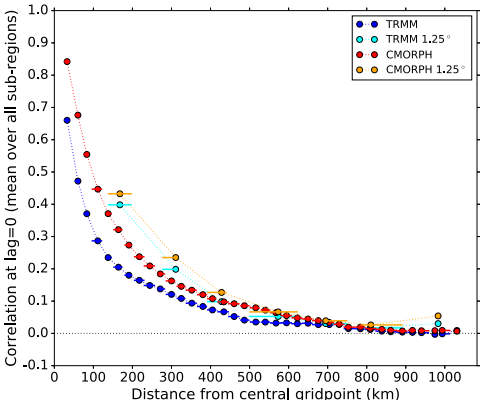


Figure 2. For CMORPH (a,c,e) 0.25° and (b,d,f) 1.25° data: (a,b) filled blocks show the 2D histogram of binned values on consecutive 3-hr steps at the same gridpoint, aggregated over all gridpoints; the dashed line shows the 1D histogram, using the right-hand axis; (c,d) the lag-0 correlation between each gridpoint in a 7×7 region and the central gridpoint (0,0), averaged over all non-overlapping 7×7 gridpoint regions in the domain; (ee,f) lagged correlations between the central gridpoint in each 7×7 region and gridpoints within each range of distance on the horizontal axis away from the central point, averaged over all 7×7 regions. In (a,b), note the logarithmic color scale; in (c–f), the printed values and filled blocks show the same data~~; in~~ . In (e,f) ~~"XXX" denotes~~ we omit the bin for points less than 0.5Δx away from the central point, as no ~~data and~~ points in these datasets fall into that bin; "centre" is the auto-correlation at the central point.

a. Correlations with distance



b. Correlations with time

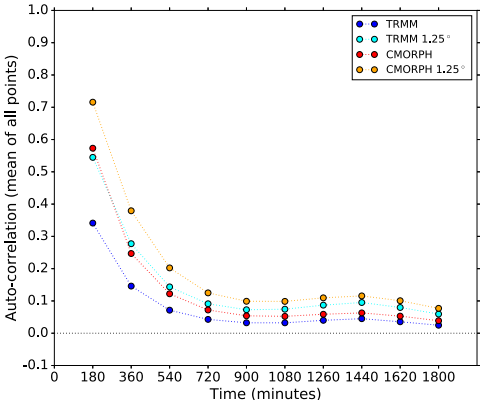


Figure 3. For TRMM 3B42 and CMORPH 0.25° and 1.25° data: (a) a measure of the spatial scale of precipitation features, computed by dividing the analysis domain into 1500×1500 km regions and calculating the lag-0 correlation between the central gridpoint and gridpoints within each distance bin (which are Δx wide, starting from 0.5Δx) away from the central gridpoint, then averaging the correlations over all regions in the domain; (b) a measure of the temporal scale of precipitation features, computed as the auto-correlation of precipitation, averaged over all points in the domain. The horizontal lines in (a) show the range of distances spanned by each distance bin; the filled circle is placed at the median distance. For clarity, we omit the correlations for zero distance and zero lag, which are 1.0 by definition.

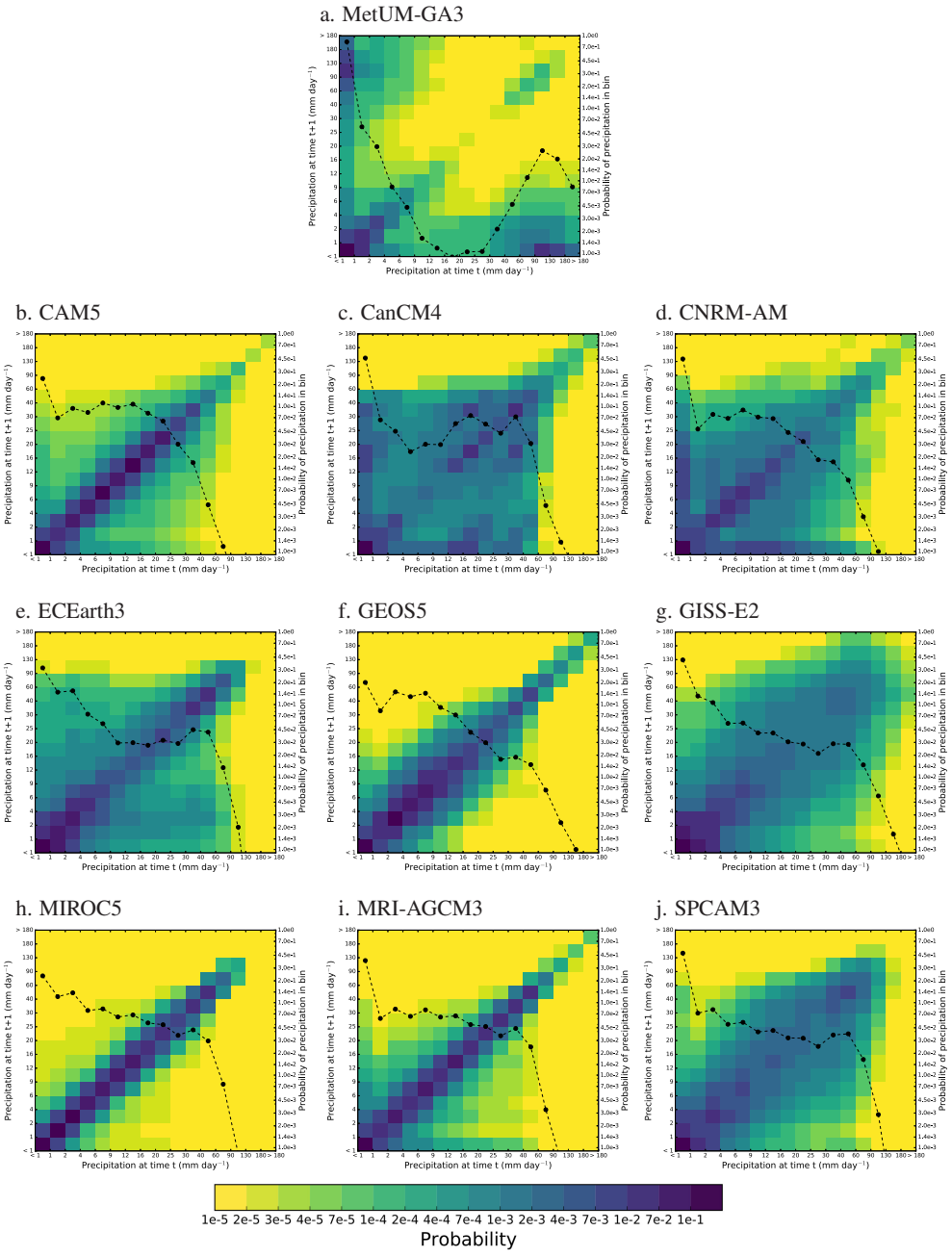


Figure 4. For each model in the GASS/YoTC dataset and using timestep precipitation data on the native model grid: filled blocks show the normalized 2D histogram of binned values on consecutive timesteps, aggregated over all gridpoints; the dashed black line shows the normalized 1D histogram, using the right-hand axis. Note the logarithmic color scale. See Table 1 for information on timestep length and grid spacing.

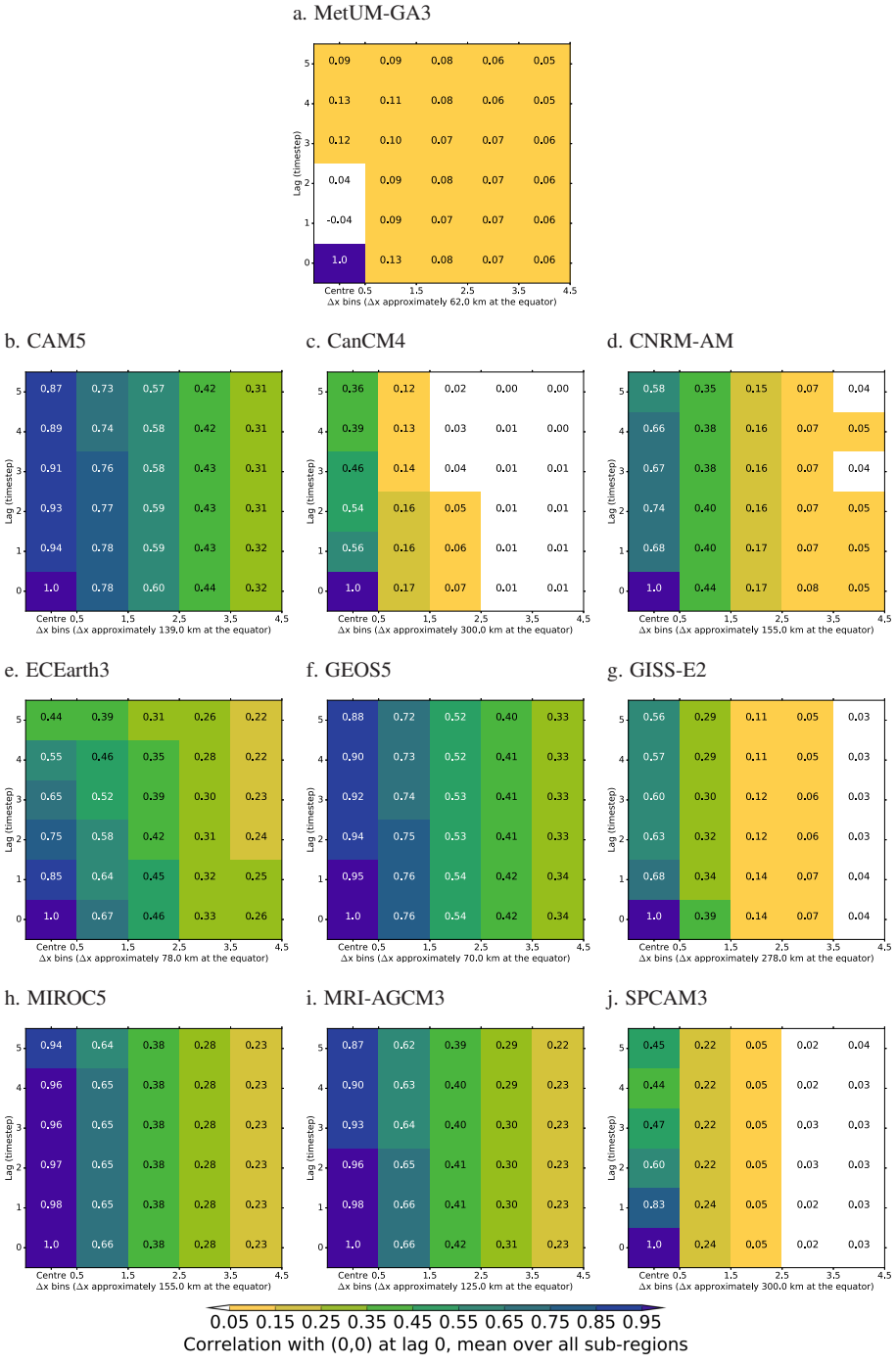


Figure 5. For each model in the GASS/YoTC dataset and using timestep precipitation data on the native model grid, lagged correlations between the central gridpoint in each 7×7 region and gridpoints within each range of distance on the horizontal axis (in units of Δx) away from the central point, averaged over all 7×7 regions. The printed values and filled blocks show the same data; "~~XXX" denotes no data;~~ "centre" in the auto-correlation at the central point. We omit the bin for points less than 0.5Δx away from the central point, as no points in these datasets fall into that bin.

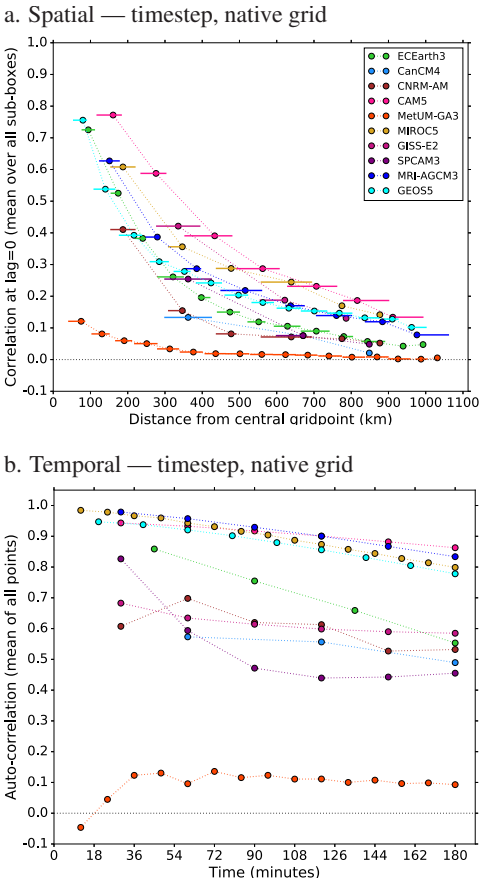


Figure 6. For the GASS/YoTC models, using timestep precipitation data on the native model grid: (a) a measure of the spatial scale of precipitation features, computed by dividing the analysis domain into 1500×1500 km regions and calculating the instantaneous linear correlation between the central gridpoint and gridpoints within each distance bin (which are Δx wide, starting from 0.5Δx) away from the central gridpoint, then averaging the correlations over all regions in the domain; (b) a measure of the temporal scale of precipitation features, computed as the auto-correlation of precipitation, averaged over all points in the domain. The horizontal lines in (a) show the range of distances spanned by each distance bin; the filled circle is placed at the median distance. For clarity, we omit the correlations for zero distance and zero lag, which are 1.0 by definition.

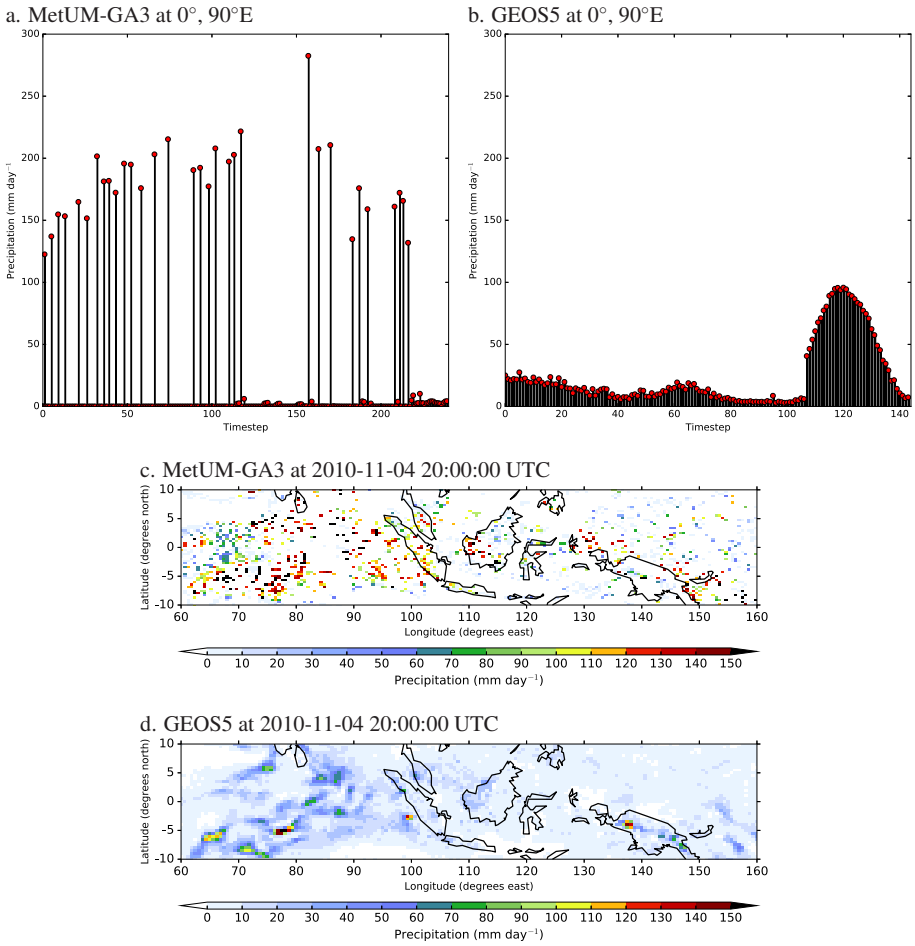


Figure 7. For (a) MetUM-GA3 and (b) GEOS5, timeseries of timestep, gridpoint precipitation (mm day^{-1}) at the point 0°, 90°E for the 48-hour forecast initialized at 00:00 UTC 4 November 2010. Each red dot represents one timestep; a red dot on the horizontal axis corresponds to zero precipitation. The series covers 48 hours for both models, which corresponds to 240 timesteps in MetUM-GA3 and 144 timesteps in GEOS5; the width of the vertical bars has been scaled for the difference in timestep length. For (c) MetUM-GA3 and (d) GEOS5, maps of instantaneous precipitation rates at the timestep corresponding to 20:00 UTC on 4 November 2010, from the forecast initialised on 00:00 UTC on 4 November 2010.

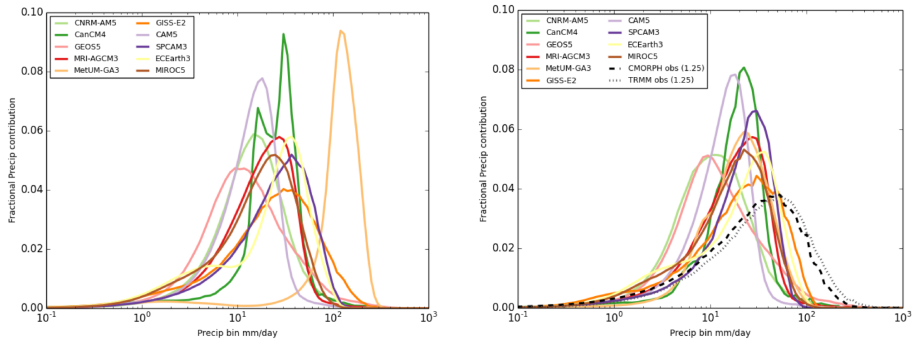


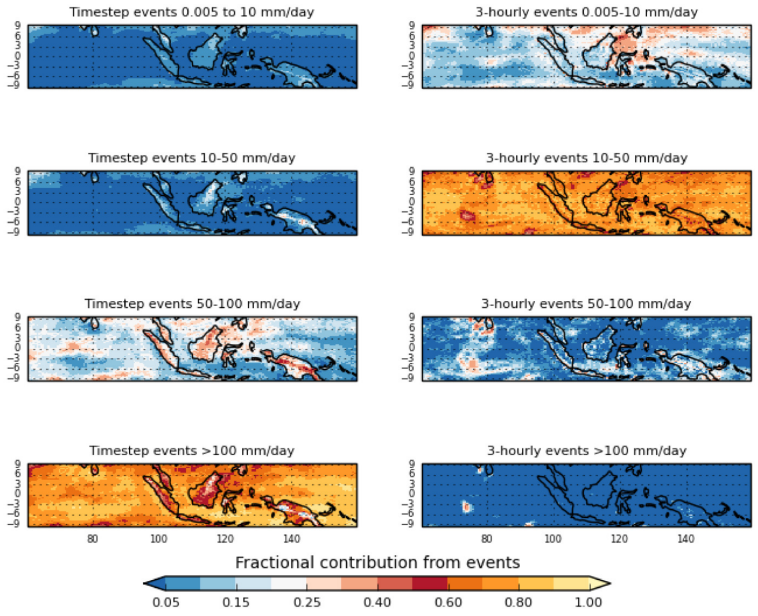
Figure 8. For the GASS/YoTC models, histograms of the fractional contributions from each precipitation bin (defined in eq. 1) to the total precipitation, computed across 60°E–160°E, 10°S–10°N, using data on the native horizontal grid and (a) timestep and (b) 3-hr averages. Panel (b) includes TRMM 3B42 and CMORPH at 1.25° resolution for the same region and time period as the GASS/YoTC models.

Table 1. For each model from the "Vertical structure and physical processes of the Madden–Julian oscillation" project from which timestep rain rates are used: the model name, the institution that produced the data, the horizontal resolution at the equator in degrees (to the nearest 0.01°) and the equivalent in km, the model timestep (Δt) in minutes and a reference with further details. Models are ordered alphabetically by abbreviation.

Model name	Abbreviation	Lon°×Lat° (km)	Δt	Reference
CAM[1]	CAM5	1.25×0.94 (139×118)	30	Neale et al. (2012)
Canadian Coupled Model	CanCM4	2.80×2.80 (311×311)	60	Merryfield et al. (2013)
CNRM-AM[2]	CNRM-AM	1.40×1.40 (155×155)	30	Voldoire et al. (2013)
European Community Model	ECEarth3	0.70×0.70 (78×78)	45	Hazeleger et al. (2012)
GEOS[3]	GEOS5	0.63×0.50 (70×55)	20	Rienecker et al. (2008)
GISS[4] GCM	GISS-E2	2.50×2.50 (278×278)	30	Schmidt et al. (2014)
Met Office Unified Model	MetUM-GA3	0.56×0.38 (62×42)	12	Walters et al. (2011)
MIROC[5]	MIROC5	1.40×1.40 (155×155)	30	Watanabe et al. (2010)
MRI[6] Atmospheric GCM	MRI-AGCM3	1.13×1.13 (125×125)	30	Yukimoto et al. (2012)
Super-Parameterized CAM	SPCAM3	2.80×2.80 (311×311)	30	Khairoutdinov et al. (2008)

[1] Community Atmospheric Model
[2] Centre National de Recherches Météorologiques Atmospheric Model
[3] Goddard Earth Observing System
[4] Goddard Institute for Space Studies
[5] Model for Interdisciplinary Research on Climate
[6] Meteorological Research Institute

a. MetUM-GA3



b. ECEarth3

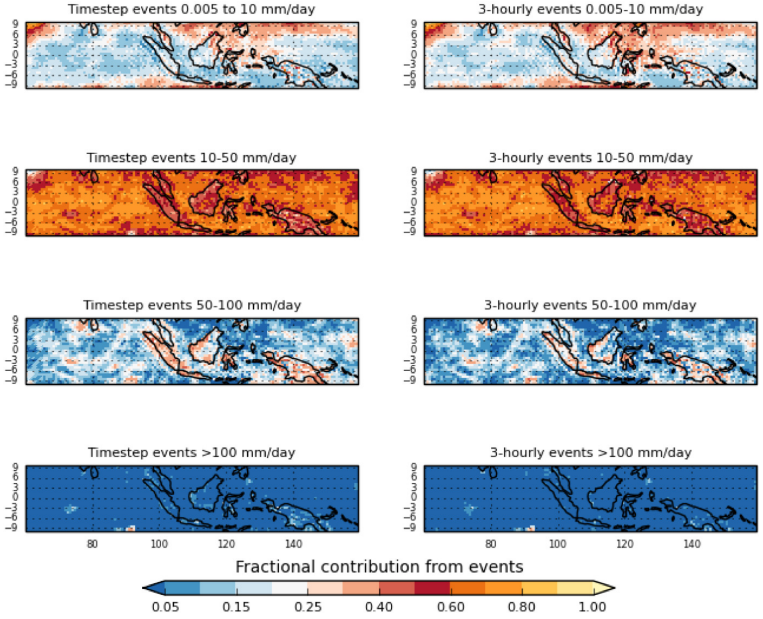


Figure 9. For (a) MetUM-GA3 and (b) ECEarth3, the fractional contributions to the average precipitation rate from ranges of intensity bins shown in the labels above each panel for (left column) timestep data and (right column) 3-hr means.

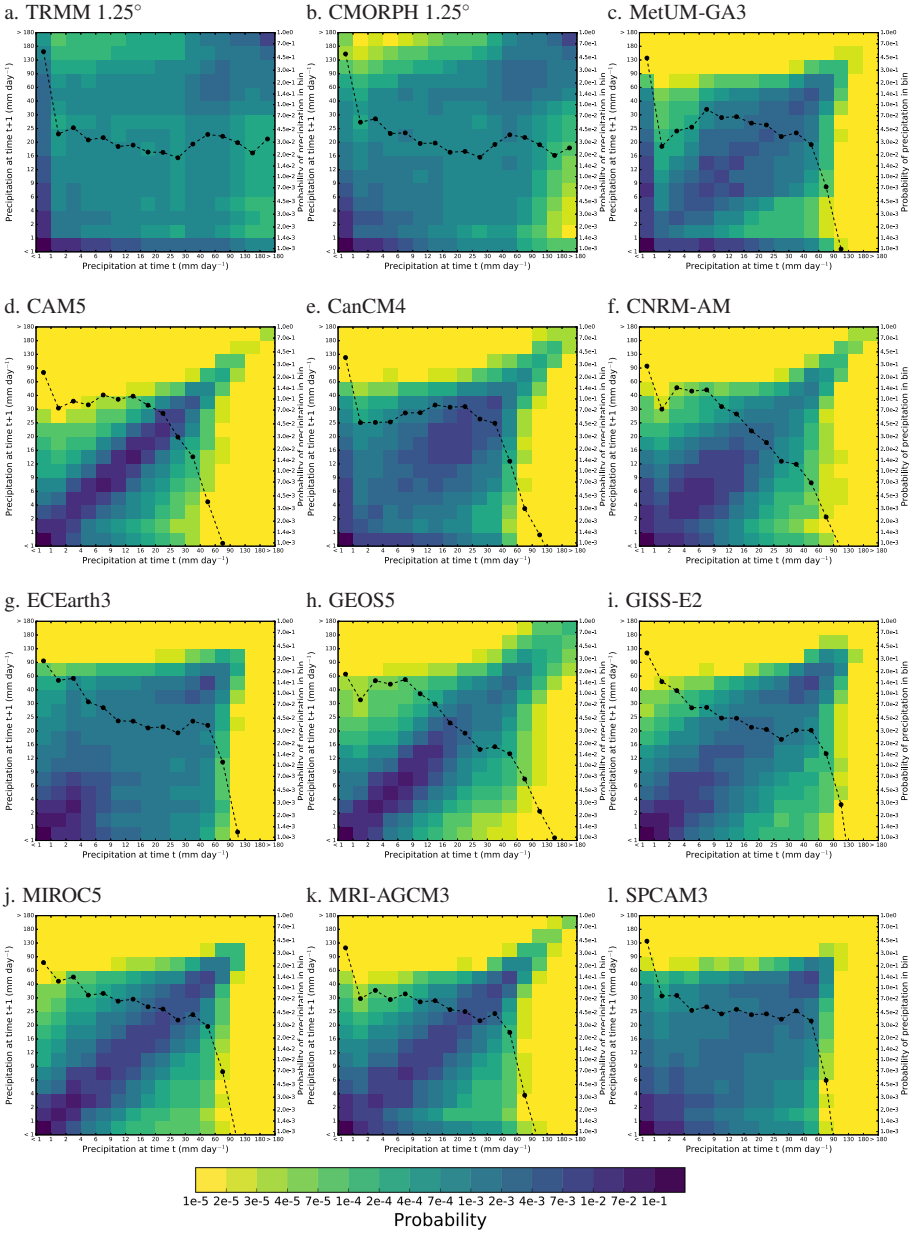
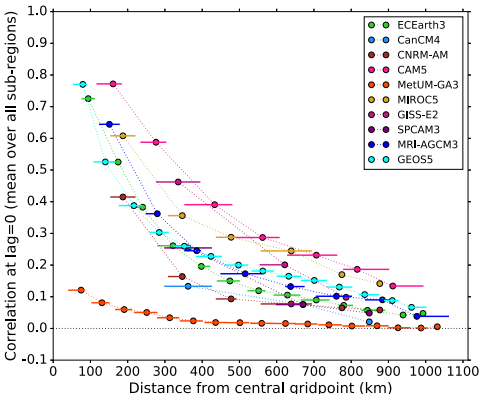


Figure 10. As in Fig. 4, but using 3-hr mean rain rates instead of timestep rain rates and with (a) TRMM 3B42 and (b) CMORPH 3-hr rain rates for the same temporal period and horizontal domain as the models.

a. Spatial — timestep, native grid



b. Spatial — 3-hr means, native grid

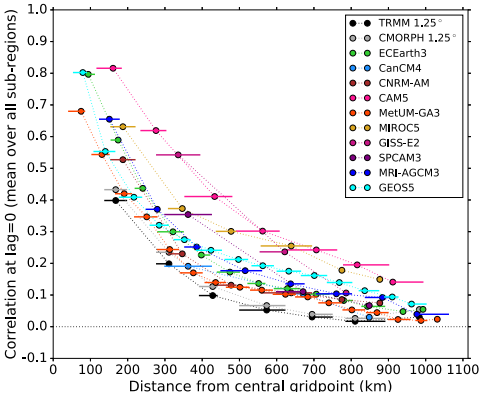


Figure 11. As in Fig. 6a, but for gridpoint precipitation data for (a) the native timestep and (b) 3-hr means. Panel (b) includes the TRMM 3B42 and CMORPH analyses at 1.25° resolution, for the same temporal period and horizontal domain as the models. Panel (a) is repeated from Fig. 6a for ease of comparison.

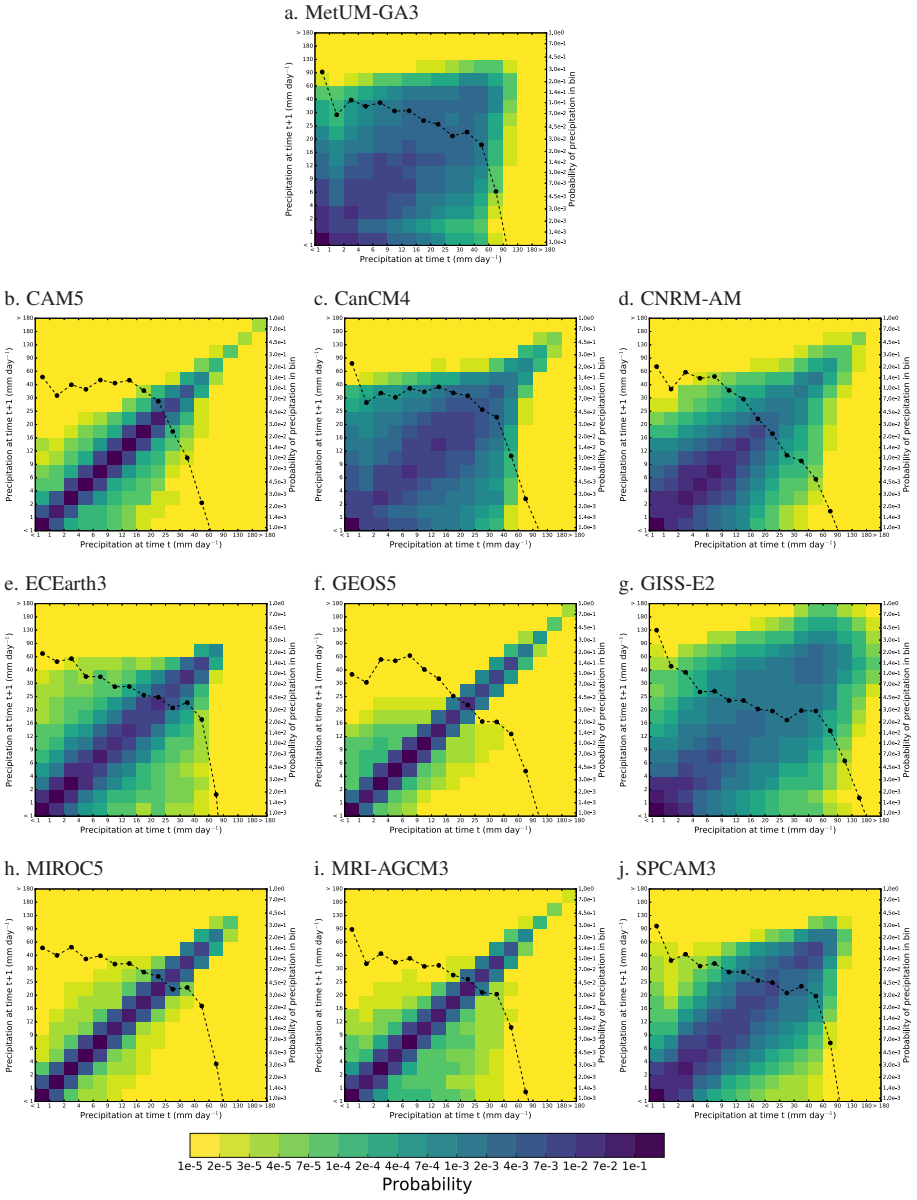


Figure 12. As in Fig. 4, but using timestep rain rates that were first spatially averaged to a $5.6^\circ \times 5.6^\circ$ horizontal grid.

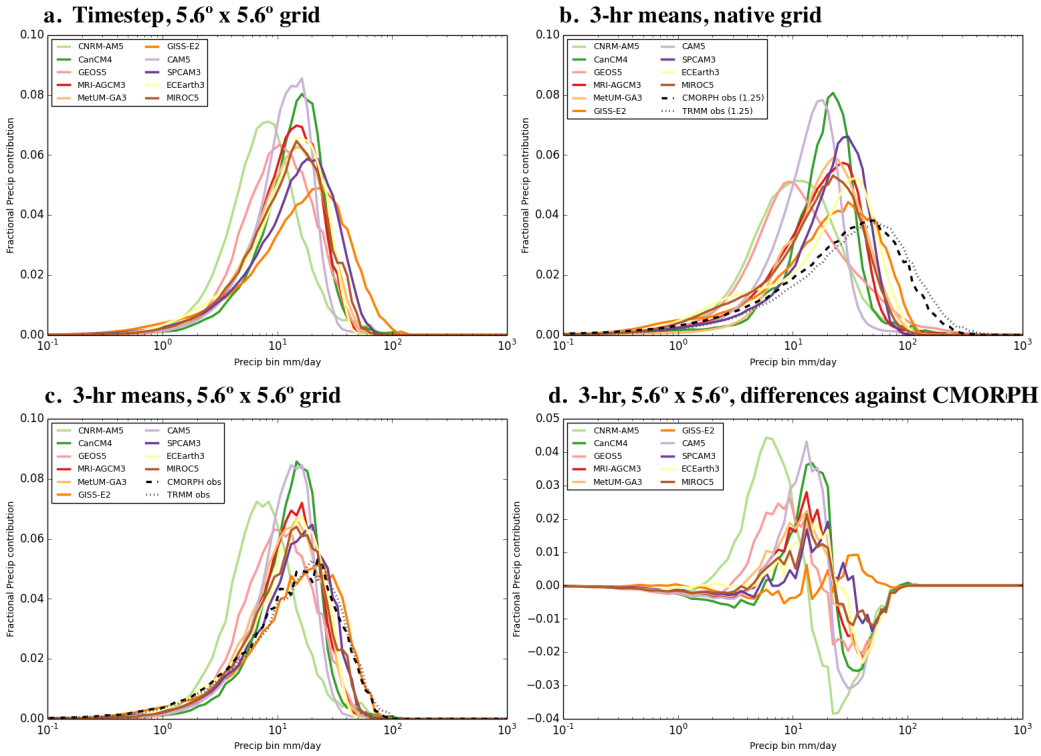
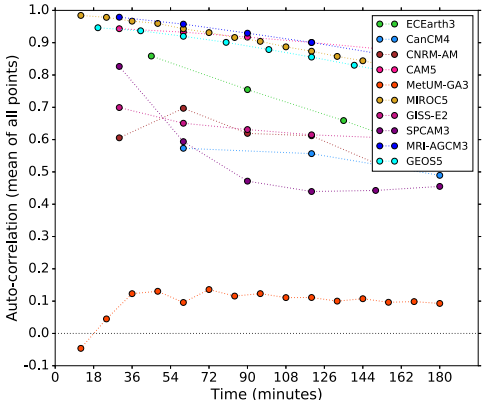


Figure 13. As in Fig. 8, but for (a) timestep precipitation rates averaged to a $5.6^\circ \times 5.6^\circ$ grid, and (b–d): 3-hr precipitation rates on (b) the native horizontal grid and (c,d) averaged to a $5.6^\circ \times 5.6^\circ$ grid. Panel (d) shows differences for each model minus CMORPH, using the $5.6^\circ \times 5.6^\circ$ data. Panel (b) is repeated from Fig. 8 for ease of comparison.

a. Temporal — timestep, native grid



b. Temporal — timestep, 5.6°×5.6° grid

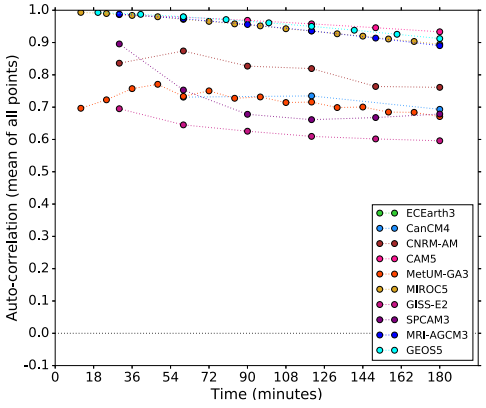


Figure 14. As in Fig. 6b, but for timestep rain rates at (a) the native gridscale and (b) averaged to a 5.6°×5.6° grid. Panel (a) is repeated from Fig. 6b for ease of comparison.

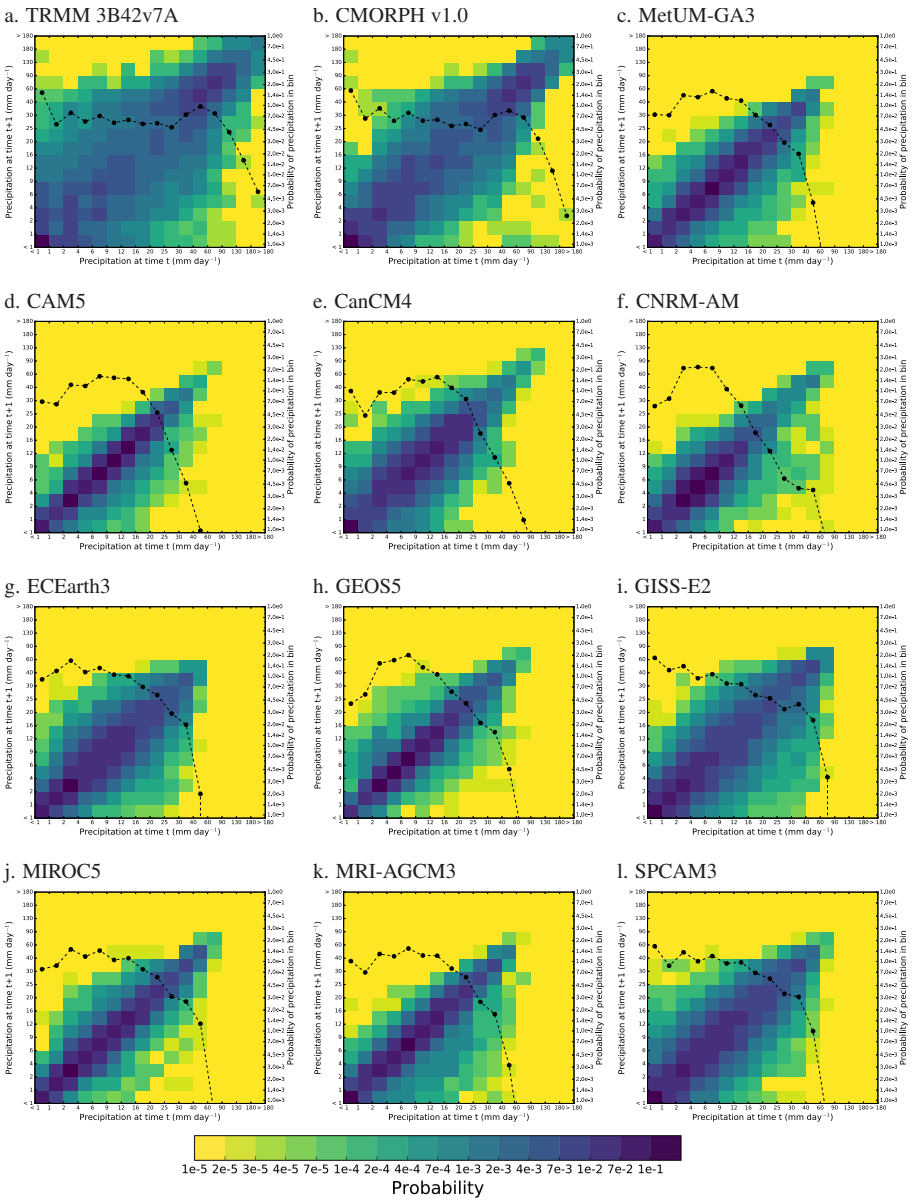


Figure 15. As in Fig. 10, but using 3-hr mean rain rates spatially averaged to a $5.6^\circ \times 5.6^\circ$ horizontal grid.

Table 2. For each model, as well as TRMM 3B42 and CMORPH: the number of timesteps in three hours; the dimensions (in native gridpoints) of the $5.6^\circ \times 5.6^\circ \times$ averaging regions discussed in the text, with the total number of native gridpoints averaged together shown in parentheses; the number of 7×7 native-gridpoint regions in the analysis domain; and the number of $\approx$1500 km $\times$ 1500 km regions in the analysis domain, with the dimensions (in native gridpoints) on each side of the region shown in parentheses. Note that while GISS-E2 and SPCAM3 have the same resolution, they have different numbers of 7×7 gridpoint and 1500×1500 km regions because of the staggering of their native grids relative to the 10°S–10°N, 60°E–160° analysis region.

Model	Δt in 3hr	Lon$\times$lat (total) in $5.6^\circ \times 5.6^\circ$	# 7×7 regions	# 1500 km (# points: lon $\times$ lat)
CAM5	6	4×6 (24)	33	7 (10×12)
CanCM4	3	2×2 (4)	5	6 (5×5)
CNRM-AM	6	4×4 (16)	20	7 (9×9)
ECEarth3	4	8×8 (64)	80	7 (19×19)
GEOS5	9	9×11 (99)	110	7 (21×27)
GISS-E2	6	2×2 (4)	6	9 (5×5)
MetUM-GA3	15	10×15 (150)	175	7 (24×35)
MIROC5	6	4×4 (16)	20	7 (9×9)
MRI-AGCM3	6	5×5 (25)	24	7 (12×12)
SPCAM3	6	2×2 (4)	5	7 (5×5)
CMORPH 0.25°	1	22×22 (484)	748	8 (55×55)
CMORPH 1.25°	1	4×4 (16)	22	8 (10×10)
TRMM 3B42 0.25°	1	22×22 (484)	748	8 (55×55)
TRMM 3B42 1.25°	1	4×4 (16)	22	8 (10×10)

Table 3. For each model, as well as TRMM 3B42 and CMORPH: summary metrics of spatial and temporal coherence in precipitation, using timestep and 3-hr data on the native horizontal grid and interpolated to a common 5.6°×5.6° horizontal grid. Positive values indicate that coherence is more common than intermittency; negative values indicate that intermittency is more common than coherence. Higher magnitudes indicate stronger coherence or intermittency for positive or negative values, respectively. The timestep column is marked "N/A" for TRMM and CMORPH because these datasets exist only as 3-hr values. By definition, the 5.6°×5.6° values are identical for the TRMM 0.25° and 1.25° datasets, as well as for the CMORPH 0.25° and 1.25° datasets.

Model	Spatial coherence				Temporal coherence			
	Native grid		5.6°×5.6° grid		Native grid		5.6°×5.6° grid	
	Timestep	3-hr	Timestep	3-hr	Timestep	3-hr	Timestep	3-hr
CAM5	0.77	0.80	0.59	0.43	0.88	0.76	0.93	0.82
CanCM4	0.23	0.28	0.47	0.30	0.38	0.57	0.46	0.66
CNRM-AM	0.41	0.52	0.41	0.35	0.44	0.53	0.59	0.71
ECEarth3	0.67	0.73	0.51	0.38	0.72	0.57	0.83	0.68
GEOS5	0.75	0.83	0.54	0.42	0.77	0.70	0.93	0.81
GISS-E2	0.54	0.55	0.57	0.45	0.69	0.68	0.68	0.69
MetUM-GA3	-0.06	0.76	0.42	0.48	0.21	0.55	0.49	0.79
MIROC5	0.65	0.67	0.61	0.48	0.92	0.71	0.95	0.81
MRI-AGCM3	0.65	0.66	0.51	0.39	0.91	0.69	0.93	0.78
SPCAM3	0.33	0.43	0.55	0.33	0.71	0.56	0.74	0.68
CMORPH 0.25°	N/A	0.80	N/A	0.34	N/A	0.41	N/A	0.73
CMORPH 1.25°	N/A	0.55	N/A	0.34	N/A	0.50	N/A	0.73
TRMM 3B42 0.25°	N/A	0.69	N/A	0.32	N/A	0.29	N/A	0.68
TRMM 3B42 1.25°	N/A	0.49	N/A	0.32	N/A	0.39	N/A	0.68