Replies to Anonymous Reviewer #1

Dear Reviewer,

Thank you for your comments and observations, which helped us to considerably improve the quality of the manuscript. We made several minor adjustments and a few major changes, which are listed here below:

- We widely revised our introduction and conclusion part - together with the section describing the stochastic physics - in order to better frame our work in the context of other existing stochastic methods in earth system modeling.
- We added two figures aimed at describing the changes in the mean state in a series of selected simulations.
- We improved the discussion of the results to better characterize the discussion of the climate variability.

You may find a detailed reply to the reviewer's comments in the following pages. We reported for completeness below each point in grey also the part of text that has been strongly modified. However, since many changes has been made extensively throughout the text, we attach to this reply also the new version of the manuscript.

Major comments:

1. **The first misleading thing is in fact the title. It needs to be changed. Instead of “in climate simulation” at the end, I suggest to use “in the ECMWF model climate simulations”.**

The reviewer is right: we therefore change the title to “*Climate SPHINX: evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth Global Climate Model*”. As highlighted in the new version of the manuscript, although EC-Earth and the ECMWF model share the main part of the code of its atmospheric model (IFS), many changes have been implemented in order to adapt it for climate simulations. Hence, calling it the “ECMWF model” may create confusion between the two.

2. **Given that all dynamical cores are perhaps equivalent to some extent, since they are based on a controlled discretization of somewhat the same governing equations, there is hope that the resolution sensitivity part of this work might extrapolate well to other climate models. However, the stochastic physics schemes considered are very particular and arguably questionable since they are essentially simply some black box random number generators that are by no means have any kind of physical justification and/or empirical validation (to the best of the reviewer's knowledge), though to their creators credit, they are the first to appear as such and led the path to stochastic parametrisations to flourish. Now days, there are many stochastic physics schemes that are developed from first principles and in most cases validated (or rather calibrated) using simulation and/or observation data, which actual achieve performances that are much better than what the pioneering ECMWF stochastic physics have achieved, in terms of improving climate simulations. To name a few, I list a couple of published articles that merit citation in this article because the readers of GMD deserve it.**

- **Ajayamohan, R. S., B. Khouider, A. J. Majda, and Q. Deng (2016), Role of stratiform heating on the organization of convection over the monsoon trough, Clim. Dyn., pp. 284 1–20, doi:10.1007/s00382-016-3033-7**
- **De La Chevroti`ere, M., Mich`ele, B. Khouider, and A. Majda (2015), Stochasticity of convection in giga-les data, Clim. Dyn., doi:doi:10.1007/s00382-015-2**
- **Khouider, B., J. Biello, and A. J. Majda (2010), A stochastic multicloud model for tropical convection, Commun. Math. Sci.**
- **Peters, K., C. Jakob, L. Davies, B. Khouider, and A. J. Majda (2013), Stochastic Behavior of Tropical Convection in Observations and a Multicloud Model, J. Atmos. Sci., 70 (11), 370 3556–3575, doi:10.1175/JAS-D-13-031.**
- **Plant, R. S., and G. C. Craig (2008), A stochastic parameterization for deepconvection based on equilibrium statistics, J. Atmos. Sci., 65, 87?105, doi: 373doi:10.1175/2007JAS2263.1**
- **Wang, Y., G. J. Zhang, and G. C. Craig (2016), Stochastic convective parameterization improving the simulation of tropical precipitation variability in the NCAR CAM5, Geophys. Res. Lett., 43 (12), 6612–6619, doi:10.1002/2016GL069818, 2016GL069818**
- **Deng, Q., B. Khouider, and A. J. Majda (2015), The MJO in a Coarse- Resolution 294 GCM with a Stochastic Multicloud Parameterization, J. Atmos. Sci., 72 (1), 55–74, 295 doi:10.1175/JAS-D-14-0120.1.**

- **Deng, Q., B. Khouider, A. J. Majda, and R. S. Ajayamohan (2016), Effect of Stratiform 297 Heating on the Planetary-Scale Organization of Tropical Convection, J. Atmos. Sci., 73 (1), 371–392, doi:10.1175/JAS-D-15-0178.1**
- **Jesse Dorrestijn, Daan T. Crommelin, A. Pier Siebesma, Harmen J. J. Jonker, and Frank Selten, 2016: Stochastic Convection Parameterization with Markov Chains in an Intermediate-Complexity GCM. *J. Atmos. Sci.,* 73, 1367–1382, doi: 10.1175/JAS-D-15-0244.1.**
- **Sakradzija, M., A. Seifert, and A. Dipankar (2016), A stochastic scale-aware parameterization of shallow cumulus convection across the convective gray zone, J. Adv. Model. Earth Syst.,8, 786–812, doi:10.1002/2016MS000634.**
- **G. A. Grell and S. R. Freitas: A scale and aerosol aware stochastic convective parameterization. Atmos. Chem. Phys., 14, 5233–5250, 2014 www.atmos-chem-phys.net/14/5233/2014/doi:10.5194/acp-14-5233-2014**
- **Frenkel, Y., A. J. Majda, and B. Khouider (2012), Using the Stochastic Multicloud Model to Improve Tropical Convective Parameterization: A Paradigm Example, J. Atmos. Sci., 69 (3), 1080–1105, doi:10.1175/JAS-D-11- 0148.1.**

Thank you for your comments. It was negligent of us not to include more references to existing and developing stochastic approaches in our manuscript: we have substantially increased the number of references to other schemes in the introduction, both when we introduce the concept of a stochastic parametrisation scheme and when we discuss previous work that demonstrates their impact on climate simulations.
Conversely, we disagree that the SPPT and SKEB schemes used here have no physical justification: we thus have included a clear explanation together with appropriate references in the section introducing the schemes.
The rephrased part of the introduction as follows:
Instead of explicitly resolving small-scale processes by increasing the resolution of climate models, a possible alternative is to use stochastic parameterisation schemes. There has been significant progress in developing stochastic schemes over the last decade, primarily for use in medium-range and seasonal ensemble forecasts (e.g. Plant and Craig, 2008; Khouider et al., 2010; Bengtsson et al., 2013; Grell and Freitas, 2013; Dorrestijn et al., 2016; Sakradzija et al., 2016; Ollinaho et al., 2016). These schemes introduce an element of randomness into physical parameterisation schemes to account for the impact of uncertain, unresolved processes on the resolved scale flow (Palmer, 2012). Stochastic schemes have been shown to improve the reliability of probabilistic forecasts on medium range and seasonal timescales, as well as improving biases in the mean state.

3. **One important point that needs to be specifically addressed is the question of what would be the intake of more carefully designed stochastic physics schemes in the context of the present study?**
This is definitely a challenging point for the whole climate community: stochastic schemes have been extensively used on shorter time scales, from forecast to decadal prediction. This is one of the first times that such schemes are implemented and tested in such a comprehensive way in a climate simulation, following the paradigm of the "seamless prediction" where the same weather forecast stochastic physics schemes are used on multidecadal timescales. Also, within Climate SPHINX we have at disposal a large number of ensemble members that are used to assess the robustness of the impact of SPPT and SKEB schemes in areas with large internal variability as the mid-latitudes.
We definitely agree with the reviewer that testing other schemes is a key question that should be addressed in the following years. In the updated version of the manuscript we thus

encourage other studies using different schemes but with a similar setup in order to provide a robust comparison also on climate time scales. Indeed, taking advantage of the reviewer suggestion of point 2 and 3, we now discuss this in the conclusion, highlighting that the resolution sensitivity can potentially be useful for all climate models, while sensitivity to SKEB and SPPT schemes cannot apply to a whole range of available stochastic schemes.

Given the similarities between the dynamical cores of climate models, since they are all based on a controlled discretization of the same governing equations, we hope that the resolution sensitivity aspect of Climate SPHINX will be useful to the whole climate modelling community. On the other hand, several promising stochastic schemes exist, and the sensitivity of EC-Earth to SPPT and SKEB described here cannot be easily extrapolated to these alternative approaches. Nevertheless, considering that Climate SPHINX is the first large experiment where stochastic schemes are used massively on the climate time range, we hope that this work paves the way for other climate-oriented simulations aimed at investigating the impact of different stochastic schemes on climate variability.

4. Even in the precise case of the present study, I was hoping to see some simulations where the two stochastic schemes (SPPT v.s. SKEB) were taking apart separately instead of considering them jointly. For instance which of these two leads to the improvements at low resolution and which one is causing the degradation at high resolution.

The reviewer is definitely right: this would be an interesting development that has not been faced within SPHINX. We widely discussed this before beginning the simulations and we concluded that it was better to avoid it for mainly three reason: 1) one of the primary goals of SPHINX is to sample the internal variability of the Earth's atmosphere, thus using the largest number of ensemble members possible. We have evidence that for mid-latitude climate even 10 members are not enough: subdividing the simulations into SPPT and SKEB cases would have affected our analysis further 2) The operational ECMWF seasonal forecasting system is currently using both the schemes at the same time. In this way we could provide an extension of their operational setup to climate simulations. 3) Finally, sensitivity tests as suggested have been carried out in a previous study by Weisheimer et al. (2014), using the same IFS model cycle in a seasonal forecast configuration (System4). These authors concluded that the SKEB schemes played a minor role for the impacts of stochastic perturbations on both the mean state of the model and the forecast quality.

However, we hope to have the chance in the upcoming SPHINX RELOADED project (a recently approved PRACE project that is just started: http://www.prace-ri.eu/projects/) to investigate this feature more in detail.

5. Page 2, Line 22: " a cheaper alternative is to use stochastic parameterisation schemes". This statement is probably true if one is ought to use physically based stochastic schemes as demonstrated in papers from the list provided above. In the case of the scheme considered here, the original purpose was simply to increase the spread in ensemble prediction runs. The idea of them (SPPT and/or SKEB) competing with high resolution models seems to be new to this article. In the reviewer's opinion, these particular schemes are not at all the best suited for the job.

Thank you for your comment. It's true that a key motivating factor behind the original implementation of SPPT and SKEB was improving the reliability of ensemble forecasts by correcting the overconfidence of the ensemble. However, many additional benefits of the schemes have been documented in weather and seasonal forecasts (e.g. Weisheimer et al, 2014). Furthermore, several recent studies indicate an improvement in the modelled climate on including the SPPT and/or SKEB schemes. For example, Dawson and Palmer (2015) demonstrate an improvement in extra-tropical weather regimes on including SPPT which is very similar to the improvement observed on increasing the resolution of the model (Dawson et al, 2012). In addition, SPPT was found to transform the representation of ENSO in coupled simulations with CCSM (Christensen et al, 2016) – the improvement is very similar to that observed when the model resolution is increased (J. Berner, pers. comm). Nevertheless, we agree that our phrasing is too strong. We have re-phrased our statement, and have stressed that the Climate SPHINX project aims to investigate the degree to which stochastic schemes can be used as a cheaper alternative to increased model resolution.

6. Page 3, Line 30: One important pre-requisite to using a stochastic parameterization under grid refinement is that the scheme has scale awareness. However, this issue is completely ignored in this paper except for some minor remark in the conclusion section on the necessity of retuning the stochastic physics at high resolution---it is unfortunate that this much computing time was wasted in brute force computations to arrive toward an obvious conclusion as this one, especially given that the literature about this issue exists (see previous list for a sample)

The SPPT scheme is in principle scale aware as it applies the perturbation to the total tendencies derived from the individual physical parametrisation schemes. So it does not require retuning with resolution. We have highlighted this when introducing the SPPT scheme.

The SPPT scheme requires no retuning with changing horizontal resolution: the same noise characteristics are used operationally at ECMWF across model resolutions. This is because the multiplicative nature of the scheme applied to the total of all parameterised physical tendencies results in perturbations to the model that scale automatically as the parametrised tendencies scale with resolution. The resolution dependence of the individual contributions from the parametrisation schemes is implicitly dealt with at the (deterministic) parametrisation level.

The SKEB scheme uses scale-dependent backscatter ratios, as discussed in the original manuscript.

However, there are indeed aspects of the deterministic model that can be retuned to compensate for a shift in the mean state introduced by the stochastic physics scheme or by a change in model resolution. We chose not to retune the model - in both its stochastic and deterministic version - as it makes it harder to compare the effect of changing resolution or including stochastic physics. What aspects of the changed model climate are due to resolution /stochasticity and what aspects are due to the tuning carried out? The same approach here used is the one that is being taken in the PRIMAVERA project and by the HiResMIP community (Haarsma et al, 2016). We have clarified this motivation in the section on tuning.

It is important to note that the tuning of the radiative fluxes has been carried out only for the T255 deterministic model version: the radiative balance has not been tuned in the higher resolution or stochastic models This ensures a clean comparison between simulations at different resolutions and with and without stochastic physics. If the model were retuned for each run, it is not possible to determine whether it is changing the tuning parameters, changing the resolution, or including stochastic physics that is responsible (Haarsma et al., in review, 2016). Including the SPPT scheme led to a negative bias in the surface heat fluxes of about 0.8 W/m_2, likely caused by a different distribution of the clouds.

7. Page 9, Lines 10-13: "unrealistic" in what sense? This is not intuitive at all as one would expect in terms of climate variability. It would probably make sense if we are dealing with an initial value problem. The fact that the member with smallest RMSE in sea-ice is chosen seems like an ad hoc choice and cannot possibly be a non biased "representative" of the ensemble. Thus the problem. What would one do if the RMSE is not accessible?

We understand that RMSE is an arbitrary measure of the spread, and other measures could be used. However, we have two main choices here: either to use the ensemble mean, or to select one single ensemble member.

Generally speaking, the mean of a bidimensional field is always unrealistic, because it is a physical state that never occurs in the daily variability. This is especially true for a field that is characterized by a binary function (SIC generally has a 0 or 1 value, with very few grid points with intermediate values) and it is has a large spatial interannual variability.

Given this consideration, choosing the ensemble mean would lead to a SIC reconstruction far to be one of possible physical solution of the system in the future, smoothed in space and with reduced interannual variability.

Therefore we are obliged to select one single ensemble member: using the RMSE measure we avoid to pick one of the outliers, selecting the ensemble member that is “closer” to the ensemble mean. We simply describe the criterion used, with no claim that other choices would not be equally suitable. We added a sentence in the text to specify that other methods can be used.

Considering that an ensemble mean would be unrealistic, especially for a field with a large spatial variance as sea ice, we select a single ensemble member representative of the ensemble. Member “r8i1p1” has been chosen to characterise the ensemble, since its climatology shows the smallest SIC Root Mean Square Error (RMSE) when compared to the ensemble mean climatology in the time window 2038-2068. Clearly, using RMSE is only one of the possible metric to perform such selection: our main goal is to pick an ensemble member that is not an outlier when compared to the other EC-Earth CMIP5 ensemble members.

Other Comments and typos:

Page 1, Line 18: “prediction of climate”. Which climate?
Corrected to “prediction of Earth’s climate”

Page 2, Line 12: “as the jet stream” ... “, the Euro Atlantic”. Insert “the” in both places.
Corrected.

Page 3, Line 8: which stochastic physics
In the new version of the manuscript we specified that we make use of the SPPT and SKEB scheme directly in the introduction. Now it reads:
Each integration is repeated with the joint implementation of two stochastic parameterizations: the Stochastically Perturbed Parameterisation Tendencies (SPPT) scheme (Palmer et al., 2009) and the Stochastic Kinetic Energy Backscatter (SKEB) scheme (Shutts, 2005; Palmer et al., 2009). In order to sample the natural variability, several ensemble members are produced for each configuration.

Page 5, Lines 8-9: Are the horizontal correlation lengths and time decorrelation parameters and standard deviations etc physically justifiable? What are the motivations behind?
A number of coarse-graining studies have provided motivation for the multiplicative nature of the noise, together with the magnitude, and spatial and temporal correlation scales. We have included a comment in the text.
The rephrased a part of the description of the stochastic physics as follows:
The magnitude of the perturbation has been motivated through coarse-graining high-resolution model simulations (Shutts and Pallarès, 2014), and a recent coarse graining study has also provided justification for the noise temporal and spatial correlation scales

Page 5, Line 22: Are the SPPT and SKEB scheme activated and deactivated simultaneously? Why not also consider cases when only one or the other is activated? (See major comment 4 above).
Yes, the two schemes are activated and deactivated together. For further details, see the reply to major comment 4.

Page 6, Sentence in Lines 27-28: How much impact does this variation in the location of the Gaussian peak has on the simulations? Are we sure we are still comparing apple to apple? What are the reason behind this difference?
The new form of the GWD parameterization is identical to that currently used by ECMWF in its operational version of IFS, while the previous one derived from an ad-hoc parameterization used only by ECMWF for its System4 model (seasonal forecast). As described in the paper, the new form has been shown to work well in recent cycles of IFS, and in EC-Earth it has been proven to be able to reproduce a realistic QBO at all resolutions considered. Overall this choice is a typical tuning choice, as many other choices which are made normally during a project development cycle. The same parameterization will enter the final version of EC-Earth 3.2 which will be used for future CMIP6 simulations.
We never use the old scheme in this project, and hence believe we are comparing similar experiments (“apple to apple”).

We now highlighted in the text that the new parametrisation is the new standard in the ECMWF forecast model while the other one was derived from the previous seasonal forecast system. Now the discussion on QBO reads as follows:

Further tuning has been performed in order to produce a realistic Quasi-Biennial Oscillation (QBO) at higher resolutions.

The EC-Earth 3.1 non-orographic gravity waves scheme is characterised by a momentum flux that is continuously launched at in the mid-troposphere to simulate the effect of gravity waves. The latitudinal profile of this momentum flux governs the correct parameterisation of gravity waves: a too high amplitude of the momentum flux will disturb the QBO in equatorial zones, particularly at high resolutions, while a too low value will lead to unrealistic eddy-driven jets, especially in the southern hemisphere, where orographically induced wave drag is low. With the current latitudinal profile, the QBO was simulated only at standard resolution (T255 with 91 vertical levels). Following advice from ECMWF staff, a resolution-dependent parameterisation of non-orographic gravity wave drag replaced the version-dependent parameterisation present in EC-Earth 3.1 (an ad-hoc parametrization developed for the ECMWF System4 seasonal forecast system). Namely, instead of using a low momentum flux average value (GFLUXLAUN=0.02) with a positive Gaussian peak at 50°S, we use a higher value (GFLUXLAUN=0.0375) which is reduced with a Gaussian shape at the equator. This negative peak is slightly deeper for stochastic runs than for deterministic simulations to compensate the effect of the stochastic noise. The average value of the momentum flux was further reduced with increasing resolution (starting from T799) according to the ECMWF specification for IFS cy40r1 (see Table 1).

Page 7, Line 5: “more”-->”more than”; “the majority of them were carried”

Thanks, corrected.

Page 7, line 20: Do all ensemble members use the same initial data?

As we explain in the text, the ensemble members are extracted the Initial Conditions are taken from the ECMWF ERA-Interim Reanalysis for 01/01/1979. A first experiment is run at each resolution for a few days, and it is then used to create the ICs for the other experiments. For instance, for T255 experiments, ICs for the 10 ensemble members are extracted using the midnight values (00:00) from each of the first 10 days respectively, and then reassigned to the 1st of January. For the T511 and T799, where we have less ensembles, only 6 and 3 days are used.

We now rephrased the text in order to make it clearer.

The Initial Conditions (ICs) in both the PDA and FSA experiments are taken from the ECMWF ERA-Interim Reanalysis (Dee et al., 2011) for 01/01/1979. A first experiment is run at each resolution for a few days, and it is used to create the ICs for the other experiments. For instance, for T255 experiments, the ICs for the 10 ensemble members are extracted using the midnight values (00:00) from each of the first 10 days respectively, and then reassigned to the 1st of January.

The same ICs are used also for FSA: in order to account for the land-surface adjustment to the new forcing, a 1-year spin up has been carried out for FSA (which is therefore starting from 2038).

Page 7, Line 31: What is the RCP8.5 scenario? Provide a reference.

We rephrased and we added the Moss et al (2010) and Taylor et al. (2012) reference.

Greenhouse gases (GHGs) and ozone concentrations as well volcanic aerosol are set according to the CMIP5 protocol (Moss et al., 2010; Taylor et al., 2012).

Page 8, Line 7: Avoid acronyms in section title, especially when they are not previously defined.

The reviewer is right: we changed the text and we added a definition of SST and SIC in the Chapter 3 in order to be able to use the acronym at that point of the manuscript

Page 9, Line 6: which reconstruction?

We added a reference to Step1Hadisst to the text.

Page 9, Line 27: explain in what sense the SST pattern is physically consistent with the sea-ice pattern.
The method used to fill the "bare points" is thought to create SSTs smoothly moving from one dataset (Step1HadISST) to the other (CMIP5 EC-Earth r8i1p1) creating a field where SSTs close to zero are found in the proximity of the sea ice cap. We improved the explanation in the text.

This provides a pattern of SSTs physically consistent with SICs: indeed, it avoids unrealistic values of SSTs in the proximity of the polar cap during winter and - using the CMIP5 EC-Earth data - it provides a reasonable distribution of SSTs in summer, where in the future scenario the ice coverage in the NH often disappears. Moreover, there is no discontinuity at the border with Step1HadiSST. The new dataset is defined as FutureHadiSST2.1.1.

Page 11, Line 11: "number ... is extremely dispersive"? Could say "Climate aspects are diverse" or "number of diverse climate aspects is large".
We rephrased this, thanks.

Page 11, Line 13: "in terms of variability rather than in terms of mean state". This can come at this expense of a hugely deteriorated mean state. Without a proper control of the model's mean state, the sole analysis of the variability by itself is not acceptable.
The reviewer is right, the presentation of the results in the previous version was not sufficient. However, we are unable to include, in the present GMD manuscript, the full description of the mean state due to page and figure limits: we have 5 different resolutions, decades of variables, stochastic physics schemes and so on. This would touch several topics and therefore makes the manuscript unnecessary long. Several of us co-authors as well as other collaborators are working on other manuscripts to follow, describing many aspects of the mean state and variability in these simulations.
Hoping to provide a reasonable equilibrium between the length of the text and the usefulness of the information, we thus decided to add a section introducing the results where we show in a 6-panel plot the sensitivity of precipitation and upper tropospheric zonal wind (200 hpa) to resolution and stochastic schemes for a selected high resolution (T799) and a low resolution version (T255).
The two figures are reported for completeness here below, while the full description and caption is now present in the new version of the manuscript.

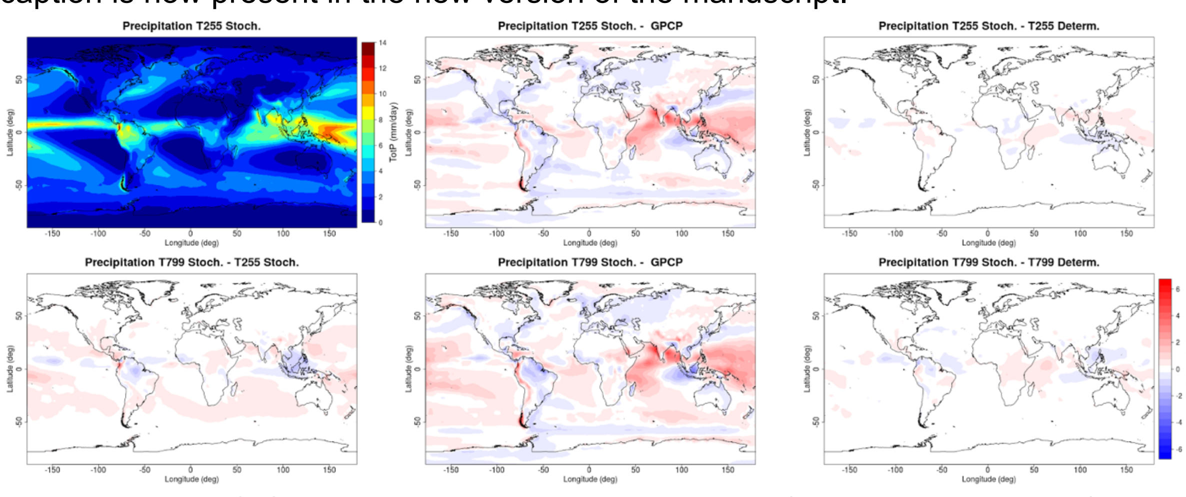


Figure R1. Upper left: Climatological ensemble mean precipitation for the PDA experiments (1979-2008) for T255 with stochastic physics. Lower left: T799 stochastic minus T255 stochastic precipitation. Central panels:

T255 (upper) and T799 (lower) precipitation bias with respect to GPCP. Right panels: Stochastic minus deterministic climatological precipitation for T255 (upper) and T799 (lower).

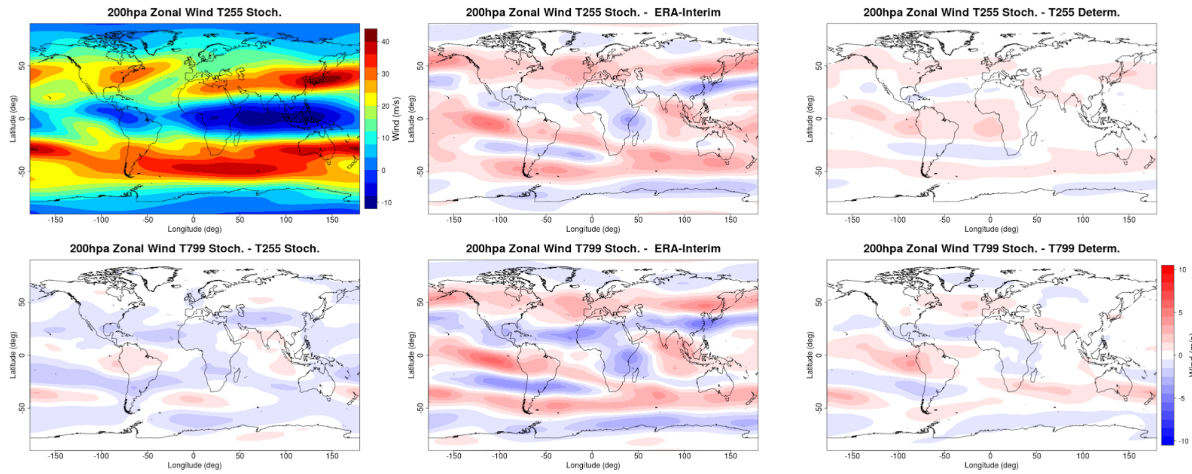


Figure R2: Same as R1 but for zonal wind at 200 hPa. Here bias is evaluated against ERA-Interim reanalysis.

The new discussion in the manuscript is here reported:

Although a detailed analysis of the mean climate in all the simulations performed would be excessively long to be included in the present work, we introduce a couple of figures showing the sensitivity to resolution and stochastic physics parameterization of the climatology of precipitation (Figure 5) and 200 hpa zonal wind (Figure 6). We compare the ensemble mean average fields of a low resolution version (T255) with a high resolution one (T799), in both its deterministic and stochastic configurations. Data have been interpolated on common 2.5°x2.5°grid.

The precipitation model bias - shown in Figure 5, with respect to Global Precipitation Climatology Project (GPCP) (Huffman et al., 2001) dataset - is especially strong in Indian Monsoon region, with an excess of precipitation that ranges from the Indian Ocean to the Western Pacific. More generally, EC-Earth tends to underestimate the precipitation over the continents and overestimate it over the oceans. When the comparison is carried out between stochastic and deterministic configurations, it is possible to see that SPPT and SKEB do neither improve nor deteriorate the climatology at both T255 and T799 resolutions.

Conversely, a slight more evident change is seen comparing the high and low resolutions: here T799 shows a widespread increase of the extratropical precipitation. But again, when it is evaluated against the model bias such changes are minor.

Impacts on the upper-tropospheric zonal wind field are clearer and they are shown in Figure 6. The T255 version - compared against ECMWF ERA-Interim reanalysis (Dee et al., 2011) - shows too strong jets in both the hemispheres. The subtropical jet over Asia and Pacific is also poleward displaced, while equatorial easterly jets are too weak. Again, stochastic physics bring minor changes, with a slightly stronger Atlantic jet, more penetrating over Europe. Conversely, the higher resolution leads to an overall weakening of the upper tropospheric winds: this is especially true over North America and the Tibetan Plateau, suggesting that this change may be induced by the stronger surface drag caused by the higher orography.

More generally, in these and other climatological fields (not shown) the impact of stochastic physics and resolution appears to be small if compared to the model bias. Indeed, larger benefits from increasing resolution and stochastic physics are expected more in terms of variability rather than in terms of mean state.

Therefore, in the following sections we will focus on a few selected features of climate variability. We will investigate the improvements and/or deteriorations following resolution increases and including stochastic parameterisation of three different phenomena: the distribution of the intensity of tropical rainfall, the tropical variability related to the Madden-Julian Oscillation and the mid-latitude variability associated with atmospheric blocking.

Figure 5, add legends in panels b) and d).

We added the legend now for both panels.

Page 11, Line 27: "occur too infrequently ... both observational datasets". This is simply not true. T799&T1279 seem to be as frequent as GPCP at for rain rates larger than 50 mm/day but much less compared to TRRM.
Figure 7b shows that for all of the deterministic model configurations, the frequency of rain rates between 20-60mm/day is below that of GPCP and TRMM. However, we think that our wording was not clear because we had not mentioned figure 7b in the text at this point, so we mentioned figure 7b earlier in the text.

Page 12, Line 16-17: How can the stochastic scheme reduce the conversion of vapour to rain. From the description in Section 2, SSPT doesn't care how the parameterization acts on the water vapour field all it does is amplify or reduce the water vapour sink or source according to the random realization of the number xi.
We agree that this sentence was not worded well and have reworded it.
One hypothesis to explain this effect is that the stochastic perturbations sometimes increase the moistening tendency of the air, so that it occurs more often that there is a high amount of water vapour in the air and heavier rain events can occur, and there is a compensating decrease in the frequency of moderate rain events.

Page 12, Line 15: In what sense Raymond et al. 2015 is an adequate reference for MJO studies.?
This reference was added to cite the importance of large-scale balanced dynamics in the Tropics and their interaction with convection, as described in Raymond et al. 2015. We have now added other MJO specific references here such as {Zhang et al. 2013; Khouider et al., 2011}.

- *Zhang, C., 2013. Madden-Julian oscillation: Bridging weather and climate. Bulletin of the American Meteorological Society, 94(12), pp.1849-1870.*
- *Khouider, B., St-Cyr, A., Majda, A.J. and Tribbia, J., 2011. The MJO and convectively coupled waves in a coarse-resolution GCM with a simple multicloud parameterization. Journal of the Atmospheric Sciences, 68(2), pp.240-264.*

Wheeler and Hendon Index is not at all the proper metric for measuring the "dynamics and thermodynamics" of the MJO. Why not do a Wheeler Kiladis spectral diagram, Hovemuller diagrams, and zonal and vertical structure. These are more more representative metrics for an MJO. The RMM index can be very misleading, see Katherine H. Straub, 2013: MJO Initiation in the Real-Time Multivariate MJO Index. *J. Climate,* 26, 1130–1151, doi: 10.1175/JCLI-D-12- 00074.1
We agree with the reviewer that Straub et al., 2013 has shown that the Wheeler Hendon index is biased towards capturing the variance in the wind signal and not as much of the OLR and convection signal of the MJO. We have now added this as a caveat in our manuscript text to interpret the results better.

We note that the Wheeler Hendon Realtime Multivariate MJO (RMM) index has been shown to be deficient in detecting MJO events when largescale circulation signals of the MJO are missing (Straub, 2013).

Yet, the Wheeler-Hendon RMM index is a standard for monitoring the state of the MJO both in climate models as well as in real time and in different operational forecasting models (Lin et al. 2009; Gottschalck et al. 2010; Rashid et al. 2011; Hamill and Kiladis 2014). Hence, we choose to use this standard index as we would like to describe the MJO characteristics in all these experimental runs with a succinct summary characteristic. We are currently working on another manuscript to study and describe the more detailed differences in the MJO in the different simulations and also for reasons of why we see the differences. For instance, we have computed the lead-lag correlation diagnostic for the precipitation and winds associated with the MJO for all the different runs and the wavenumber-frequency spectral diagrams. Fig. R3 shows an example for the T255 experiment for one ensemble member showing that the SPPT simulation at this low resolution tends to improve the correlation of the propagating precipitation signal past the Maritime continent. We have also computed the power spectra for the winds and OLR in the Tropics for the different experiments and see a similar enhanced power in the intraseasonal band for winds (shown in Fig R4) and OLR in the runs with SPPT active.

- Lin, H., G. Brunet, and J. Derome, 2009: An observed connection between the North Atlantic oscillation and the Madden–Julian oscillation. J. Climate, 22, 364–380, doi:10.1175/ 2008JCLI2515.1
- Gottschalck, J., and Coauthors, 2010: A framework for assessing operational Madden–Julian oscillation forecasts: A CLIVAR MJO working group project. Bull. Amer. Meteor. Soc., 91, 1247–1258, doi:10.1175/2010BAMS2816.1.
- Rashid, H. A., H. H. Hendon, M. C. Wheeler, and O. Alves, 2011: Prediction of the Madden–Julian oscillation with the POAMA dynamical prediction system. Climate Dyn., 36, 649–661, doi:10.1007/s00382-010-0754-x
- Hamill, T. M., and G. N. Kiladis, 2014: Skill of the MJO and Northern Hemisphere blocking in GEFS medium-range reforecasts. Mon. Wea. Rev., 142, 868–885, doi:10.1175/ MWR-D-13-00199.1.

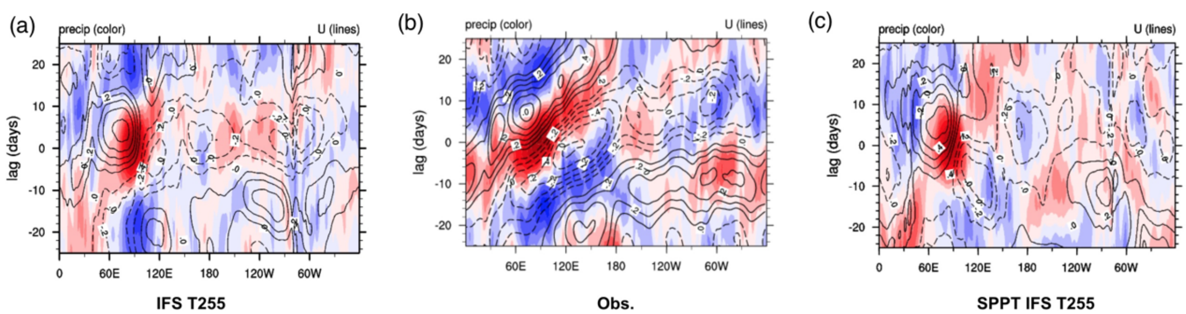
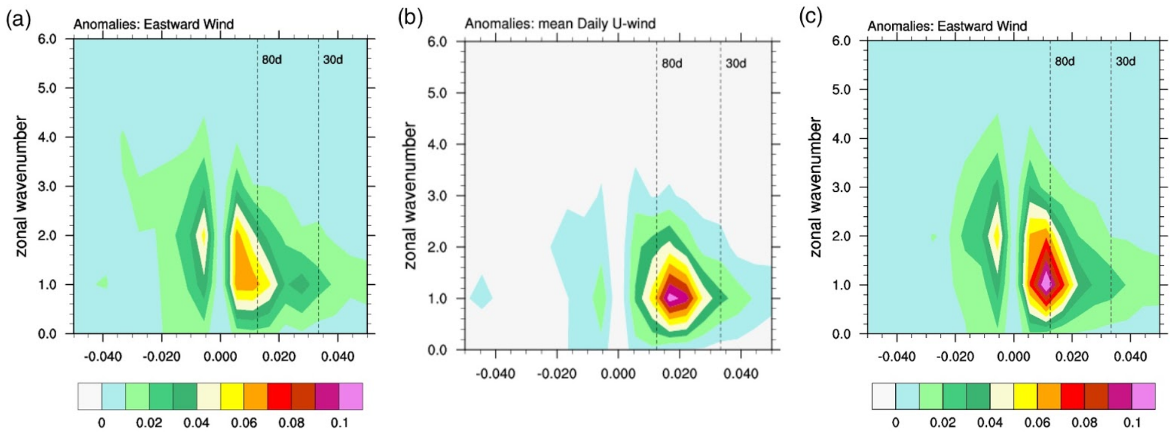


Fig R3: Lag–longitude diagram (a) EC-Earth (b) GPCP precipitation and NCEP winds and (c) SPPT IFS of 10°N–10°S-averaged intraseasonal precipitation anomalies (colors) and intraseasonal 850-hPa zonal wind anomalies (contours) correlated against intraseasonal precipitation at a Indian Ocean reference region. The two IFS simulations were done at T255 resolution. Contours and colors are plotted every 0.1. The zero line is not shown.



IFS Obs SPPT IFS

Fig R4: November-April wavenumber-frequency spectra of 10°N-10°S averaged a) EC-Earth 850 hPa U winds, b) NCEP 850 hPa zonal wind and c) SPPT IFS 850 hPa zonal wind. Units for the zonal wind spectrum are m^2 s^{-2} per frequency interval per wavenumber interval.

Replies to Anonymous Reviewer #2

Dear Reviewer,

Thank you for your comments and observations, which helped us to considerably improve the quality of the manuscript. We made several minor adjustments and a few major changes, which are listed here below:

- We widely revised our introduction and conclusion part - together with the section describing the stochastic physics - in order to better frame our work in the context of other existing stochastic methods in earth system modeling.
- We added two figures aimed at describing the changes in the mean state in a series of selected simulations.
- We improved the discussion of the results to better characterize the discussion of the climate variability.

You may find a detailed reply to the reviewer’s comments in the following pages. We reported for completeness below each point in grey also the part of text that has been strongly modified. However, since many changes has been made extensively throughout the text, we attach to this reply also the new version of the manuscript.

Comments:
In the following I will address how the evaluation section could be improved and extended.

0. I would have liked to see some mean field comparison of temperature, wind, precipitation and in particular their variability.
The reviewer is right, the presentation of the results in the previous version was not sufficient. However, we must keep in mind that we cannot include in the present GMD manuscript the full description of the mean state: we have 5 different resolutions, decades of variables, stochastic physics schemes and so on. This would touch several topics and therefore makes the manuscript unnecessary long.
Hoping to provide a reasonable equilibrium between the length of the text and the usefulness of the information, we thus decided to add a section introducing the results where we show in a 6-panel plot the sensitivity of precipitation and upper tropospheric zonal wind (200 hpa) to resolution and stochastic schemes for a selected high resolution (T799) and a low resolution version (T255).
The two figures are reported for completeness here below, while the full description is now present in the new version of the manuscript.

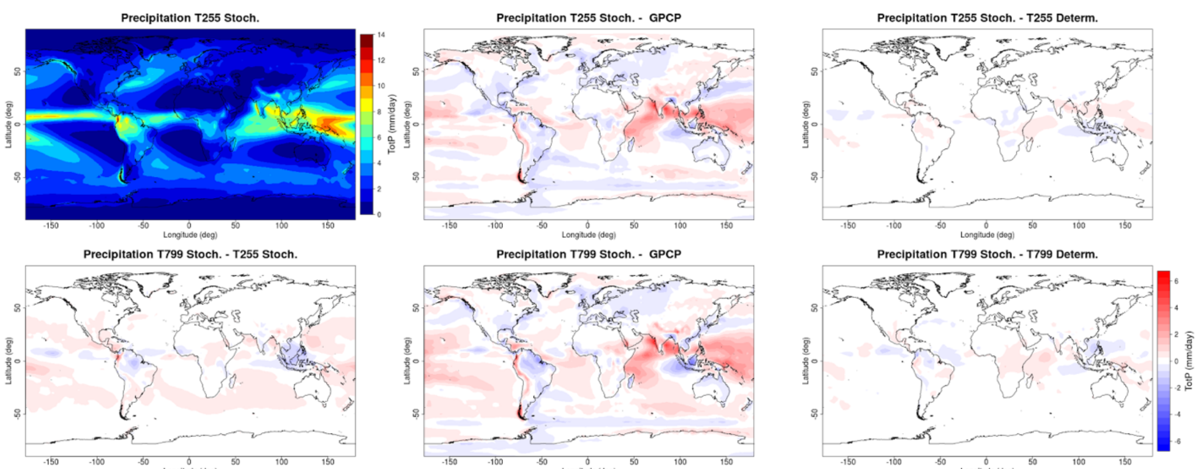


Figure R1. Upper left: Climatological ensemble mean precipitation for the PDA experiments (1979-2008) for T255 with stochastic physics. Lower left: T799 stochastic minus T255 stochastic precipitation. Central panels: T255 (upper) and T799 (lower) precipitation bias with respect to GPCP. Right panels: Stochastic minus deterministic climatological precipitation for T255 (upper) and T799 (lower).

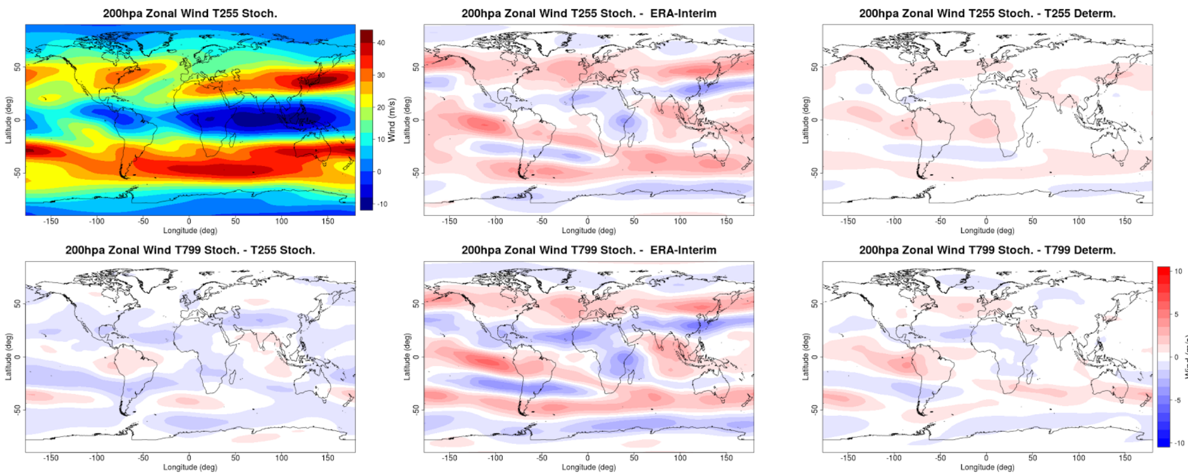


Figure R2: Same as R1 but for zonal wind at 200 hPa. Here bias is evaluated against ERA-Interim reanalysis.

The new discussion in the manuscript is here reported:

Although a detailed analysis of the mean climate in all the simulations performed would be excessively long to be included in the present work, we introduce a couple of figures showing the sensitivity to resolution and stochastic physics parameterization of the climatology of precipitation (Figure 5) and 200 hpa zonal wind (Figure 6). We compare the ensemble mean average fields of a low resolution version (T255) with a high resolution one (T799), in both its deterministic and stochastic configurations. Data have been interpolated on common 2.5°x2.5°grid.

The precipitation model bias - shown in Figure 5, with respect to Global Precipitation Climatology Project (GPCP) (Huffman et al., 2001) dataset - is especially strong in Indian Monsoon region, with an excess of precipitation that ranges from the Indian Ocean to the Western Pacific. More generally, EC-Earth tends to underestimate the precipitation over the continents and overestimate it over the oceans. When the comparison is carried out between stochastic and deterministic configurations, it is possible to see that SPPT and SKEB do neither improve nor deteriorate the climatology at both T255 and T799 resolutions.

Conversely, a slight more evident change is seen comparing the high and low resolutions: here T799 shows a widespread increase of the extratropical precipitation. But again, when it is evaluated against the model bias such changes are minor.

Impacts on the upper-tropospheric zonal wind field are clearer and they are shown in Figure 6. The T255 version - compared against ECMWF ERA-Interim reanalysis (Dee et al., 2011) - shows too strong jets in both the hemispheres. The subtropical jet over Asia and Pacific is also poleward displaced, while equatorial easterly jets are too weak. Again, stochastic physics bring minor changes, with a slightly stronger Atlantic jet, more penetrating over Europe. Conversely, the higher resolution leads to an overall weakening of the upper tropospheric winds: this is especially true over North America and the Tibetan Plateau, suggesting that this change may be induced by the stronger surface drag caused by the higher orography.

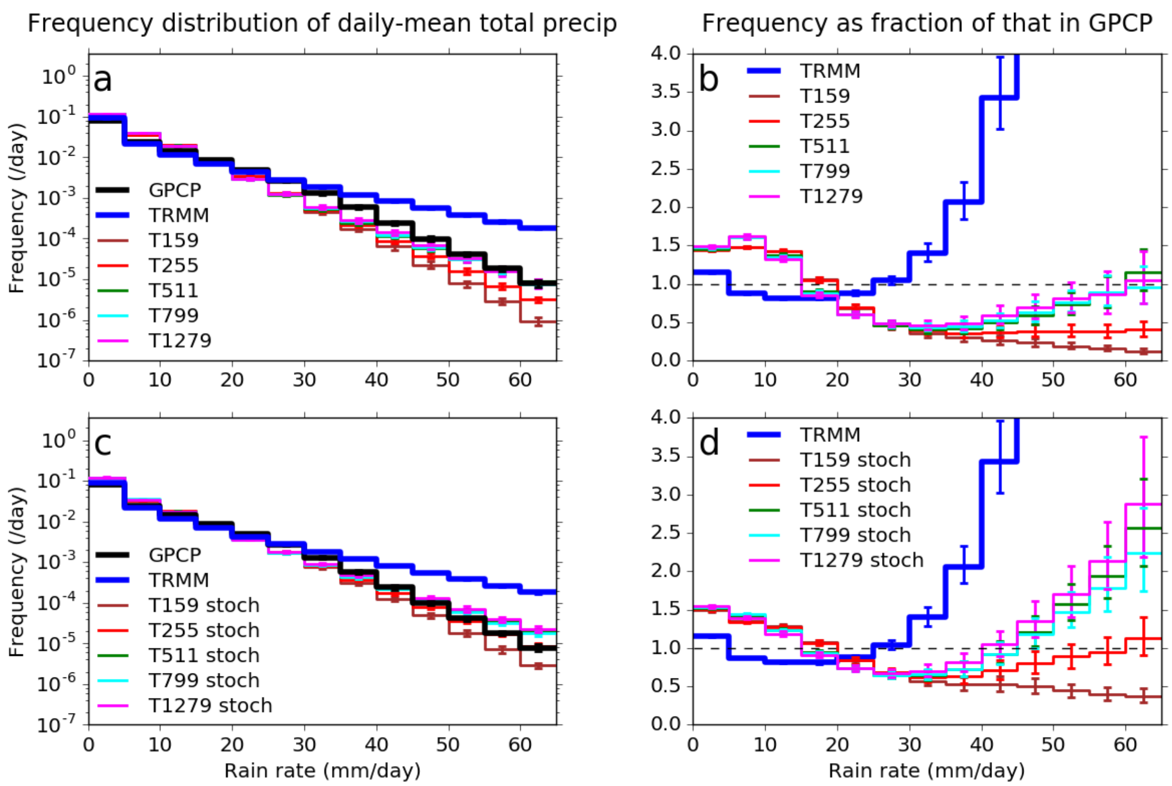
More generally, in these and other climatological fields (not shown) the impact of stochastic physics and resolution appears to be small if compared to the model bias. Indeed, larger benefits from increasing resolution and stochastic physics are expected more in terms of variability rather than in terms of mean state.

Therefore, in the following sections we will focus on a few selected features of climate variability. We will investigate the improvements and/or deteriorations following resolution increases and including stochastic parameterisation of three different phenomena: the distribution of the intensity of tropical rainfall, the tropical variability related to the Madden-Julian Oscillation and the mid-latitude variability associated with atmospheric blocking.

Tropical Rain:

1. The manuscript interprets the difference between the TRMM and GPCP datasets as observational uncertainty. The TRMM data should be added to the right panels displaying the frequency fractions with regard to GPCP.

We added in the Figure 5 TRMM data too. The new figures are reported here below.



2. Since in the observational datasets the differences in the rainfall rates of 40mm/day are larger than the differences between the model simulations, a crude significance test would help with the interpretation of the results. This could be as simple as displaying rainfall rates for the first and second half of the simulation period.
In order to evaluate the uncertainties, we calculated 95% confidence intervals using a bootstrap method. These confidence intervals have been added on in Figure 5 to show the sampling uncertainty of the frequencies. The method is explained in the text:
Vertical bars in figure 7 show the 95% confidence intervals of the frequencies associated with sampling uncertainty. These were calculated using a bootstrap method. For each dataset, a surrogate dataset was created by randomly sampling individual years of data with replacement. The frequency distribution of the surrogates and their frequency ratios with respect to the GPCP surrogate were calculated. This was repeated 1000 times to produce the distribution of the calculations associated with sampling uncertainty, from which the confidence intervals were derived

3. I would suggest a second figure showing of rainfall rates and frequency fractions for the small rainfall rates only, so that the differences in the curves can be seen better.
We appreciate the suggestion, but in our view this would lead to there being too many figures, making the paper excessively long. This is especially true considering the new figures on the mean state. We have made sure that important details of the results for low rain rates are spelt out in the text.

MJO
4. Figure 6: Please specify how the patterns for the 4 phases of the MJO are computed. Are they based on all data, or are they calculated for the observations only? It seems the comparison of the frequency of occurrence and the amplitude of the MJO should

be done in a common basis system. Otherwise a case must be made, why different basis system are used and the patterns for each experiments (or at least some) shown and the change in patterns discussed.

Yes, the MJO index was computed for the period 1980-2001. We have mentioned this in the manuscript now. The observations are also for same period. The MJO is essentially a mode of internal variability in the atmospheric system and hence, our MJO analysis in this study is mainly focused on statistical differences between the model MJO variability and the observational variability based on the same period of evaluation.

Figure 8 shows the frequency of occurrence vs. the mean amplitude of the MJO in the four different regions around the Tropics for all the different runs (colours) and for ECMWF ERA-Interim Reanalysis (grey, Dee et al., 2011) over the 1980-2001 period.

5. Please refer to Weisheimer et al. 2014, who report that in particular SPPT improves MJO in System 4.

We thank the reviewer for this comment. We have now included a reference to the Weisheimer et al. 2014 result suggesting that the SPPT scheme helps improve MJO representation in the seasonal forecasting system.

6. Are these results significant? Again a dot for each first half and second half of the simulations would be sufficient.

We have now included error bars in this figure by computing the same statistic for periods half the length of the original analysis as suggested.

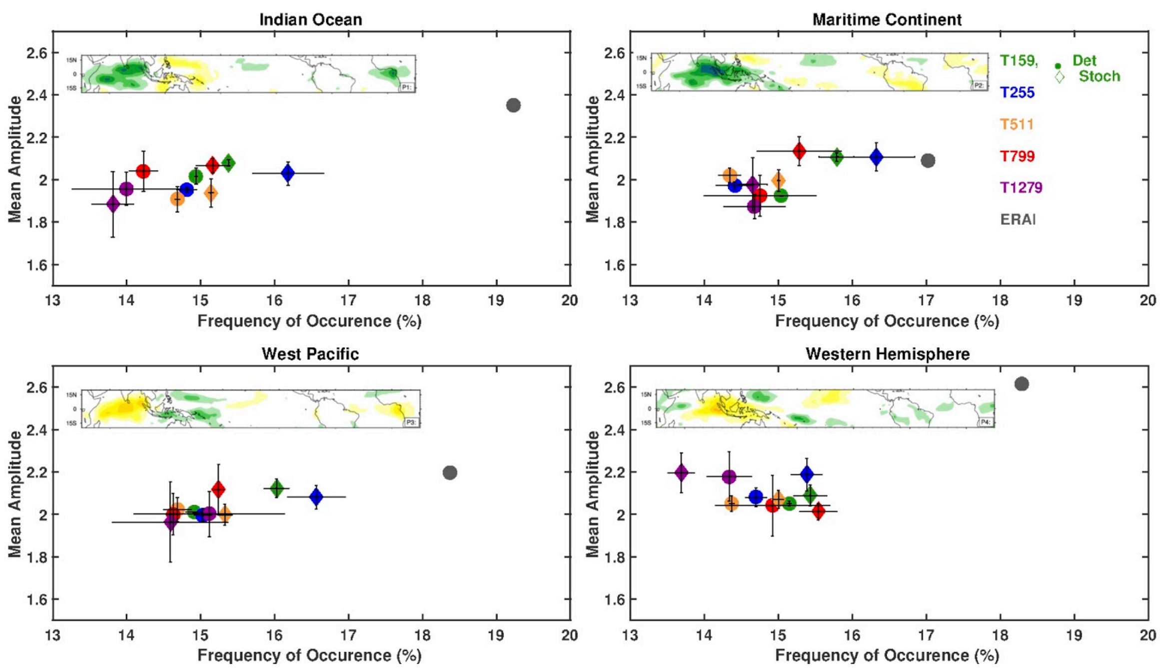


Figure R3: New version of MJO figure (new figure 8) from the manuscript.,

Blocking

7. Dawson et al. 2015 is in the Bibliography, but I couldn't find a reference in the text. This paper should be clearly cited in the blocking section.

Dawson and Palmer (2015) has been cited in the introduction. In addition, since as stated by the reviewer is relevant for the discussion on atmospheric blocking, it is cited also in the blocking and conclusion sections.

8. Other work has suggested a positive impact of stochastic physics on blocking. Why is this not the case here? Maybe it only helps at horizontal resolutions of less than T159? Is there really no benefit even at resolutions of T159?

It is hard to say it without a direct comparison of the diagnostics used: at this stage we can only provide some hypothesis.

A first reason may be exactly as stated by the reviewer, i.e. stochastic physics providing benefits for blocking at lower resolution than T159.

A second one could be related to the metrics adopted: for instance, blocking diagnosed from weather regimes (Dawson and Palmer, 2015) is not always associated with blocking as evaluated directly from blocking indices. In this sense, for instance, the instantaneous blocking index here reported does not take into account the duration of the events.

The last possibility can be deducted from Figure R4 here attached. It shows the full behaviour of the blocking diagnostic in all the ensemble members from Climate SPHINX. Dashed lines are the stochastic runs and solid ones are the deterministic ones. Bold lines mark the ensemble mean, suggesting that the two versions are extremely similar, indistinguishable from a statistical point of view. Indeed, the natural variability is extremely evident in this diagnostic. It is easy to find a pair of members where a deterministic run outperforms a stochastic one or vice versa.

It may be possible that some of the studies (e.g. Berner et al., 2012, Dawson and Palmer, 2015) reporting an improvement of blocking with stochastic schemes may be affected by sampling of the internal variability: one of the things that Climate SPHINX clearly highlights is that a very large number of ensemble members is needed if we want to draw any conclusion on the sensitivity of mid-latitude climate to the adoption of stochastic schemes.

We mentioned that in the manuscript now, including a comment in the blocking section and in the conclusions.

Conversely, in the mid-latitudes, where atmospheric blocking frequencies were analysed, no statistical difference is found between stochastic and deterministic runs. Previous works (Berner et al., 2012; Dawson and Palmer, 2015) suggest that block- ing regimes can benefit from stochastic schemes. We note that the simulations presented here are at a higher resolution than Berner et al. (2012) and have been analysed using a different metric to Dawson and Palmer (2015). Nevertheless, we must observe that our blocking diagnostic shows strong variability in the different ensemble members, suggesting that a single realization may be not enough to capture the real sensitivity of this diagnostic to stochastic schemes. On the other hand, we found that increased horizontal resolution seems extremely important to decrease the blocking bias: in agreement with other recent works (Davini and D’Andrea, 2016; Schiemann et al., 2016) this is true especially over the Euro-Atlantic sector - where the T799 resolution (~25 km) reduces it to negligible values - but not evident over the Pacific.

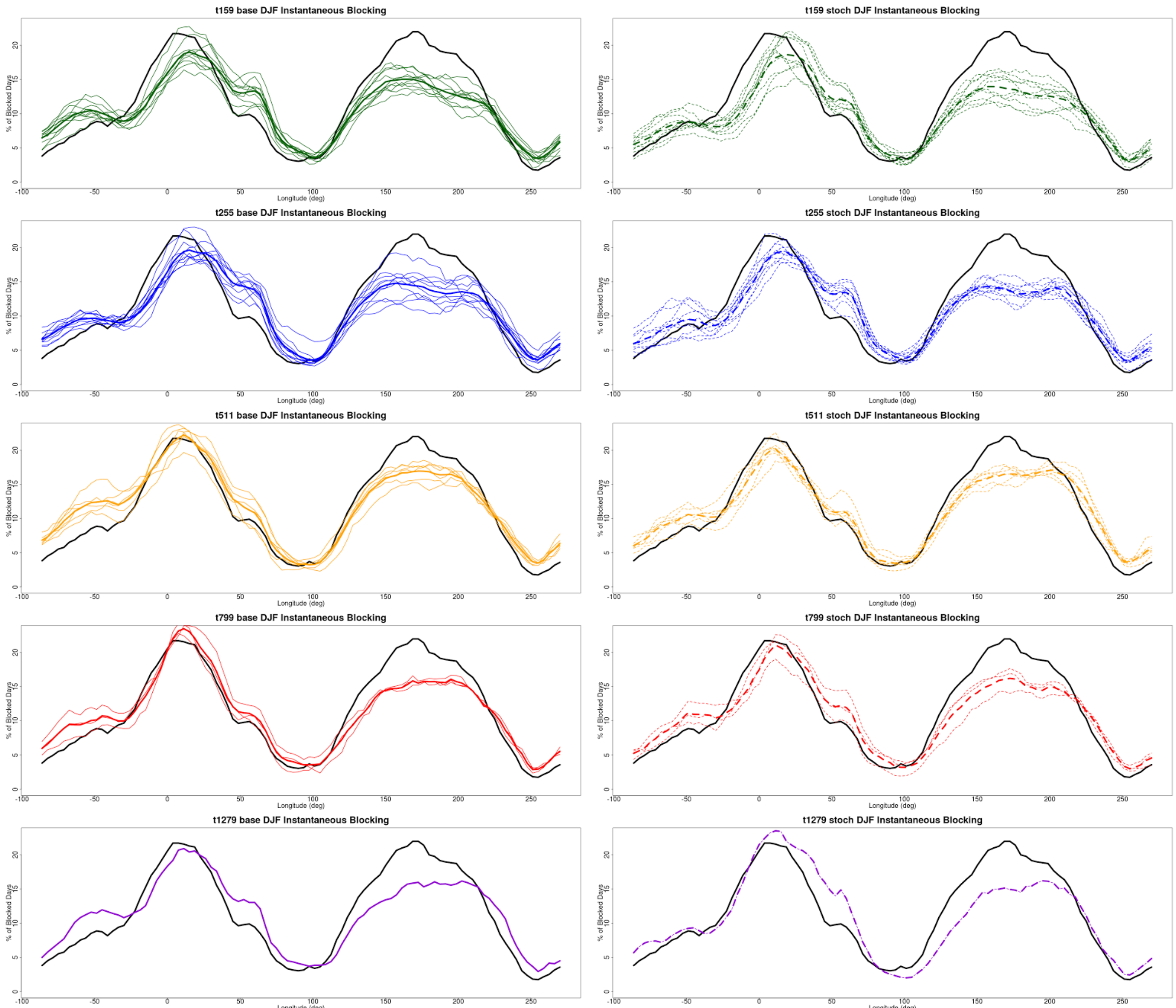


Figure R4: Winter locking climatologies for all the PDA simulations. Dashed lines are stochastic simulations, solid ones deterministic ones. Bold lines are ensemble means. Black line is the Reanalysis.