Report Test of Cycled Grell-Freitas Convective Parameterization

Global Model Test Bed (GMTB)

June 2017

Points of Contact: Ligia Bernardet (Ligia.Bernardet@noaa.gov) and Grant Firl (grantf@ucar.edu)

Additional GMTB staff that contributed to this activity: Jimy Dudhia, Michelle Harrold, Judy Henderson, Hongli Jiang, Jamie Wolff, Man Zhang.

Georg Grell contributed to the section *Developer's Comments.*

Table of Contents

List of Tables and Figures

Executive Summary

The GMTB conducted an assessment of the GFS using the Grell-Freitas (GF) cumulus parameterization (GFS-GF) and compared the results against a control run using the GFS FY17 that employs the scale-aware Simplified Arakawa Schubert (SAS) cumulus scheme (GFS-SAS). This test was a follow up to a previous test conducted by GMTB using the GFS FY16, in which GF runs were initialized from the operational GFS analyses (*cold starts*). In contrast, for the current test, cycled data assimilation was employed to produce initial conditions consistent with the model physics. Additionally, a set of GF cold-started runs was produced using the FY17 GFS (GFS-GFcold) for an assessment of the impact of cycling.

The test, which employed the GMTB Single Column Model (SCM), and the Global Spectral Model (GSM) was planned jointly by the GF scheme developer, GMTB, EMC, and representatives of NOAA's Next-Generation Global Prediction System program, with the goal of supporting the development of an advanced physics suite for the GFS. Global forecasts were run at medium resolution (T574, approximately 34 km) without tuning of the physics suite. While the GFS-GF performed better than GFS-SAS in a few metrics, such as temperature bias and precipitation frequency bias in the Northern Hemisphere, overall results indicate that the GFS-GF did not produce better forecasts than GFS-SAS, and should not progress to higher tiers of testing unless it is revised and goes through initial tuning. Since NGGPS is funding an effort to contrast, compare and potentially unify the SAS and GF schemes, it is recommended to defer further testing of GF by GMTB until results from this other effort are available. A perspective on tuning, provided by the developer, Georg Grell, has been included after the discussion section. Key findings from the test follow and are further substantiated and discussed in the body of the report.

Key Findings from Single Column Model (SCM_KF)

SCM_KF1. For the strongly forced maritime case, the GFS-GF suite produces weaker convective tendencies and convective transport than GFS-SAS. This alters the relationship among the physics schemes within the suite, leading to the explicit microphysics scheme in GFS-GF to show a greater relative response to the forcing.

SCM_KF2. For the relatively weakly forced continental convection case, the convective tendencies produced by the GFS-GF suite were generally comparable to or greater than those produced by the GFS-SAS suite.

SCM_KF3. Use of the GFS-GF suite leads to higher moisture content in the boundary layer and generally produces a higher cloud fraction throughout the column, particularly in the lower-to-mid troposphere.

SCM_KF4. During the suppressed convection phase of the maritime convective case and two subperiods of the continental convective case, the GFS-GF suite alters the interaction with the PBL scheme, leading to the transport of PBL moisture higher in the column and occasionally spuriously large cloud fraction at the PBL top.

SCM_KF5. Although both suites produce approximately the same precipitation amounts for both cases, the GFS-GF suite produces a much lower convective precipitation ratio and lower temporal variability than the GFS-SAS suite.

SCM_KF6. During the maritime deep convective period, the forcing ensemble elicits greater variability from the GFS-GF suite than the GFS-SAS suite.

Key Findings from Global Diagnostics (GD_KF) of Global Single Case

GD_KF1. GFS-GF produces extra precipitation in the tropics, especially between 5°S and 5°N.

GD_KF2. Total precipitation, and its partition between convective and explicit components, is different between GFS-SAS and GFS-GF. Precipitation development occurs faster for the cycled runs. Comparing GFS-SAS with GFS-GF, the spin up time is shorter in GFS-SAS. Compared to CMORPH observations, model precipitation in GFS-GF is too light and frequent in rainfall intensity between 2-7 mm d^{-1}.

GD_KF3. GFS-GF has more low clouds in the SH and Tropics over the ocean, which leads to a substantially different radiation budget.

***GD_KF4*.** The terms of the water budget are different between the GFS-SAS and GFS-GF, with the GFS-GF displaying higher precipitable water.

Key Findings from Global Verification (GV_KF) of Retrospective Runs

***GV_KF1*.** There is little difference in results between the cold and cycled runs with GF.

GV_KF2: For most variables and forecast lead times, regardless of global sub-region, GFS-SAS has less RMSE than GFS-GF. The fewest number of differences are noted in the SH, while the most are seen in the TROP region.

GV_KF3: The profiles of temperature bias are different between the GFS-SAS and GFS-GF with the better performer depending strongly on sub-region; the GFS-GF is preferred over the NH and GFS-SAS is generally preferred for the TROP.

GV_KF4. GFS-SAS is warmer than GFS-GF over the CONUS at 2m, and the two configurations have distinct diurnal cycle of errors: GFS-SAS warms up too quickly in the daytime, while GFS-GF maximum temperatures are below observed. A problem noted in a previous GMTB test using the FY16 GFS Physics suite, of CONUS 2-m temperatures increasing with forecast lead time in GFS-SAS runs, has not been seen in this test.

GV_KF5: Wind biases are similar between the GFS-SAS and GFS-GF throughout the atmosphere in the NH and SH, but the GFS-GF has larger negative biases in the TROP sub-region, especially at upper levels.

GV_KF6. Precipitation placement is better in GFS-SAS than GFS-GF.

GV_KF7. The configuration that predicts better precipitation coverage depends on domain. Both configurations see an increase in precipitation coverage with forecast lead time.

GV_KF8. GFS-GF is more cyclogenetic and produces more tropical cyclogenesis false alarms than GFS-SAS.

Introduction

To inform the development of an advanced physics suite for NOAA's GFS, the Global Model Test Bed (GMTB) conducted a second test of the Grell and Freitas (GF, 2014) convective parameterization. This work built on a previous test conducted by GMTB (see website and report) in two important ways. First, for relevance to transition to operations, the entire GFS system was upgraded from the FY16 to the FY17 configuration, and therefore the control (dubbed GFS-SAS) employed the scale-aware Simplified Arakawa Schubert (SAS; Han et al. 2017). Second, the experiments used additional tiers of the GMTB hierarchical testbed, including the GMTB Single-Column Model (SCM), a case study with in-depth diagnostic, and a longer period in which the experiments were run in cold start (GFS-GFcold) and cycled Data Assimilation (DA; GFS-SAS and GFS-GF) mode. Consistent with the concept of hierarchical testing, the global experiments were run using middle-resolution. This was done to save computational resources, while producing information that can be used to potentially advance the innovation to more sophisticated tiers of testing, including high-resolution global runs. Given the limited computational resources available to GMTB and the demand posed by running the 80-member ensemble as part of the NCEP Global Data Assimilation System (GDAS), the test period for global run was restricted to two weeks (June 1-15, 2016), a length of time considered minimal but sufficient to provide information about the performance of GF in cycled mode.

The plan for this test was devised jointly by the main developer (Georg Grell of NOAA ESRL/GSD), EMC, GMTB, the NOAA Next-Generation Global Prediction System (NGGPS) Physics Team co-leads, and the NGGPS Program Office. This parameterization was selected for testing because of its potential for improving forecasts. It is a state-of-the-art scheme that has been developed recently, but follows a long line of parameterizations from the developers. It includes a scale-aware feature, which make the scheme suitable for use in a wide range of model resolutions. Additionally, it incorporates an ensemble approach to the representation of convection, which can improve the forecast by using a collection of parameters and algorithms to represent the convective triggers, vertical mass flux, and closures. The ensembles can also be perturbed by stochastic fields for deterministic forecasting as well as ensemble data assimilation. Flux form tracer transport, wet scavenging, and aerosol awareness are also options in this scheme. An additional factor that led to this choice was the scheme's maturity, its history of operational use at NCEP in the RAP, and the fact that its development is funded by NGGPS. It should be noted that no tuning was performed for the GF in the context of the GFS physics suite or of the Global Spectral Model (GSM) dynamic core.

This report focuses on the experiment configuration, key findings, and discussion of results. For further information not covered in this report, please see the comprehensive verification results here.

Experiment Configuration

Single Column Model

Overview

A previous test conducted by GMTB used the GMTB SCM to test the replacement of the deep and shallow convection schemes in the 2016 GFS operational physics suite with those created by Grell and Freitas. In that test, the procedure followed that which is described in Randall et al. (2003) and Zhang et al. (2016), namely that initial conditions and column forcing are derived from observations obtained during Intense Observation Periods (IOP). The atmospheric physics suite that makes up the SCM is allowed to respond to the forcing by generating parameterized clouds and precipitation, radiative heating, vertical mixing, etc. Physics suite performance can be gauged by comparing diagnosed physical quantities to observations. While the previous test used one relatively strongly-forced deep convective case over an oceanic surface, the current test adds a case that features more weakly-forced deep convection over land. As with tests using the global model, this test uses the updated 2017 operational GFS physics as the control and the experimental configuration uses the same suite with the shallow and deep convection schemes switched to the Grell-Freitas equivalents. As with the previous test, interpretation of results from a SCM can be considered to be straight-forward due to the lack of three-dimensional interaction and feedback of error through the dynamical core, yet necessarily considered incomplete for the same reason. A physics suite's performance within a SCM context is dependent on the quality of the prescribed initial conditions and forcing, and is but one indicator of how a suite may perform within a fully-complex global model.

Source Code

The code for running the SCM portion of the test resides in NOAA VLab under the "gmtb-scm" project name. Within this Git project, the specific code for running this test can be found in the master branch under the tag v1.3. This repository contains both the GMTB SCM code and the identical branch of GFS used for the global portion of this test, specifically the branch located on EMC's SVN server at /gfs/branches/GMTB/gf_da_test. The GMTB SCM code interfaces with the GFS physics through the version of nuopc_physics.f90 found in the specified branch. Both the control runs and the experimental runs using the Grell-Freitas convection use a version of *gbphys.f* that was modified to call the Grell-Freitas scheme. The only difference between the two runs is the specification of the convection scheme, which is controlled through the variables *imfdeepcnv* and *imfshalcnv*. Further details of the GFS source that was used can be found in the source code description of the global model section in this report.

Cases

The SCM was configured to run two cases from the GCSS Fourth Working Group that focuses on deep convection: the sixth intercomparison case based on the ARM's Tropical Warm Pool - International Field Experiment (TWP-ICE) field campaign as described in Davies et al. (2013) (as done for the previous test) and the third intercomparison case based on data from ARM's Southern Great Plains (SGP) observation network in the summer of 1997 as described by Cederwall et al. (2000) and Xie et al. (2002). The TWP-ICE case is based on a suite of observations obtained near Darwin, Australia in January and

February of 2006. Meteorological conditions observed included deep convection associated with an active phase of the monsoon and suppressed convection and clear sky associated with the inactive phase. The initial profiles of temperature, moisture, and horizontal winds reflect average conditions over the study area (centered on 12.425°S, 130.891°E) at 0300 UTC on January 19, 2006. The surface is oceanic with a fixed SST, implying interactive surface fluxes calculated by a surface-layer scheme, and an observed ozone profile is included for use with interactive radiation. The effect of large-scale advection on the temperature and moisture profiles is calculated using two separate terms following the "horizontal advective forcing" method of Randall and Cripe (1999): prescribed horizontal advective tendencies plus a vertical advective term that combines the prescribed vertical velocity and the modeled temperature and moisture profiles. Horizontal wind profiles are relaxed to observed profiles on a timescale of two hours. Forcing for the SCM is supplied for the entire length of the TWP-ICE field campaign from 0300 UTC on January 17, 2006 to 2100 UTC on February 12, 2006. Following Davies et al. (2013), the simulation period was split into two time periods for analysis -- one that featured active, deep convection (from 0000 UTC on 20 January to 1200 UTC on 25 January) and one that featured suppressed, shallow convection (from 0000 UTC on 28 January to 1200 UTC on 2 February). In addition to a "best estimate" forcing dataset for the time period, a 100-member forcing ensemble is utilized to gauge sensitivity to the supplied forcing. The forcing ensemble was created by quantifying uncertainty in the surface rainfall measurement and using the constrained variational analysis method to derive 100 equally likely forcing profiles. The greatest change among the forcing datasets is in the prescribed vertical velocity, which is very sensitive to surface precipitation. The forcing ensemble is described in detail by Davies et al. (2013).

The ARM SGP case is based on a suite of observations obtained over northern Oklahoma and southern Kansas in June and July of 1997. The meteorological conditions observed over the domain include dormant, hot, and dry days as well as the passage of strongly-precipitating MCSs that are typical of the southern great plains of North America in the summer. Three separate subperiods with multiple convective events occurring inside the observation domain were identified for analysis. Subperiod A (from 2300 UTC on June 26 through 2300 UTC on June 30) captured isolated storms and the passage of a strong MCS, whereas subperiods B (from 2300 UTC on July 7 through 2300 UTC on July 12) and C (from 2300 UTC on July 12 through 2300 UTC on July 17) featured weaker forcing from isolated storms and "glancing blows" from passing MCSs. Following the case setup procedures outlined by Cederwall et al. (2000), and unlike the TWP-ICE case, subperiods are treated as separate model runs with separate initial conditions and forcings to combat model state drift. For each subperiod, the initial conditions represent the averaged atmospheric state over the observational domain at the initial time. Also unlike the TWP-ICE case, the surface is considered to be over land, but the surface fluxes are specified from observations rather than calculated by the surface scheme. In addition, the effect of large-scale advection on the temperature and moisture profiles are applied using the so-called "revealed forcing" method of Randall and Cripe (1999) -- both horizontal and vertical advective tendencies are specified and there is no feedback on the SCM forcing due to the modeled state. Horizontal wind profiles are relaxed to observed profiles according to a timescale determined by the observed wind profile and the observation domain length scale.

When observations are available for a particular quantity, a skill score was calculated following Taylor (2001) as:

$$S = \frac{4(1+R)^4}{\left(\sigma_f + \frac{1}{\sigma_f}\right)^2 (1+R_0)^4}$$

where R is the correlation coefficient between simulated and observed quantities and σ_f is the ratio of modeled to observed variance. This score is printed in the legend for each suite.

Global Model

Overview

This test expanded the scope of the previous Grell-Freitas test by adopting cycled DA for both the control and the experimental cumulus parameterizations. In this test, the two model configurations had their own independent initial conditions, both generated by running the NCEP GDAS. A third experiment configuration consisted of non-cycled GF runs, using the GFS-SAS analysis files as initial conditions, to help understand the differences between cycled and non-cycled GF results. In addition to employing cycled DA, the GMTB workflow migrated to be based entirely on the Rocoto Workflow Management System, following the major changes implemented by EMC for use in their GFS parallel runs. As a first approach to testing GFS-GF, and to conserve GMTB's limited computational resources while providing sufficient information, the model was run at a coarser resolution (T574) than the one used for the operational GFS (T1534). The results from the global model runs were investigated by looking at diagnostics from single case initializations as well as aggregated verification results in order to provide a holistic approach to understanding the differences between GFS-SAS and GFS-GF.

GFS, NEMS, GSM, and Physics

The source code for this experiment was based on the GFS code undergoing final tests for the July 2017 (Q3FY17) operational implementation at NCEP, with a modification to add the GF parameterization. The provenance of scripts and source codes is described in detail below. All revision numbers were recorded when the test started.

Code management for this test was done using the EMC Apache Subversion (SVN) server. All runs were performed using the NEMS-based GSM model. While the initial GF test performed by GMTB set *use_nuopc=true,* the cycled GF test had to employ *use_nuopc=false* due to incompatibilities when cycling was enabled. This issue was reported back to EMC. The code bases for both the global runs and the SCM cases were conducted from the branch located on EMC's SVN server at /gfs/branches/GMTB/gf_da_test (see Fig. 1 for code base schematic). The GSM code used for the global runs is under a branch created by the GMTB on March 15, 2017 from the GSM top of trunk, revision 89613, and contains identical code to the trunk and to the GSM undergoing final testing for use the planned operational implementation in July 2017 (https://svnemc.ncep.noaa.gov/projects/gsm/tags/gsm_q3fy2017_kappa), with exception to the GF code provided by the developer that has been added and integrated. Files module_*cu_gf_driver.f*, *module_cu_gf_deep.f*, *and module_cu_gf_sh.f* were added and changes were made to three files in the

GSM code: *phys/gbphys.f*, *phys/compns_physics.f90*, and *phys/gloopr.f90* to accommodate the requirements of the GF scheme, and to create the ability to select between the scale-aware SAS and GF in the runs. Swapping the scale-aware SAS scheme for the GF scheme was the only change to the operational GFS physics suite (Table 1).

In addition to GSM, the NEMSGSM application requires two other components, NEMS and Chem. As described in Fig. 1, for this test the tags for the Q3FY17 GFS were employed: https://svnemc.ncep.noaa.gov/projects/nems/tags/nems_q3fy2017_kappa and https://svnemc.ncep.noaa.gov/projects/aerosol/chem/tags/chem_q3fy2017_kappa. The GMTB conducted a test to ascertain that forecasts using scale-ware SAS were bitwise identical to forecasts created using the branch with the GF code or the trunk.

Table 1. Description of the GFS-SAS and GFS-GF physics suites.

Physical Process	Operational Suite	Experimental Suite
Convection (deep and shallow)	Scale-Aware SAS	GF
Turbulent transport (PBL)	Hybrid Eddy-Diffusivity Mass-Flux	
Radiation	RRTM for General Circulation Models (RRTMG)	
Gravity wave drag	Orographic and stationary convective	
Land surface model	Noah	
Cloud microphysics	Zhao-Carr	

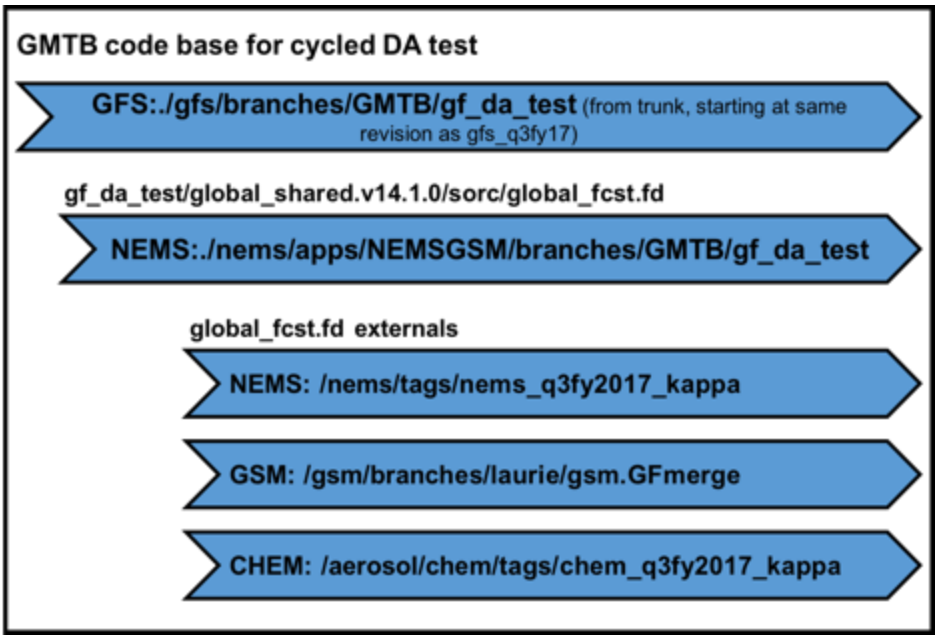


Figure 1. Schematic of GMTB code base for cycled DA test.

When the GF scheme is activated in the namelist by setting *imfdeepcnv* (deep convection; option 3) and *imfshalcnv* (shallow convection; option 3), a number of other parameters specific to GF can be set. The specific parameters used in this test are the same as the previous non-cycled Grell-Freitas test and were selected in collaboration with the developer through an iterative process of running several cases to exercise the different options (Table 2). Mid-level convection (*imid*; option 0) was turned off for this experiment. The average of all possible closures was chosen for deep convection (*ichoice*; option 0), and for the closures for shallow convection (*ichoice_s*), option 2 was selected. The diurnal cycle adjustment was also activated for this test (*dicycle*; option 1). For the GFS-SAS, the 2017 operational settings were used (*imfdeepcnv* and *imfshalcnv* set to 2*)*.

Table 2. Description of parameters used in the cycled GF Test.

	imfdeepcnv	imfshalcnv	imid	ichoice	ichoice_s	dicycle
Scale-Aware SAS	2	2	--	--	--	--
GF	3	3	0	0	2	1

Post-processing, Graphics, and Diagnostics

The *unipost* program within NCEP's UPP v7.3.2 was used to output the necessary variables at specified levels, derive additional meteorological fields, and vertically interpolate fields to isobaric levels. The post-processed forecast files included two- and three-dimensional fields, which are necessary for the

plotting routines, calculation of diagnostics, and verification tools. The necessary parameter files for *unipost* were based on those being utilized at NCEP for parallel testing. Output from *unipost* were in Gridded Binary version 2 (GRIB2) format, and the *wgrib2* utility was used to interpolate the post-processed files to a 0.25° global grid (G193).

The cyclone tracker used with the operational GFS was applied to extract information about location and intensity of storms identified by NHC and JTWC, as well as to identify tropical cyclogenesis events.

Graphics of model output from UPP were created using Python and included a suite of figures by ingesting the 0.25° GRIB2 files, and either plotting the gridded data directly, or regridding it to various verification grids used by NCEP. The test plan provides a comprehensive list of the variables plotted for each model forecast.

In addition, the GMTB also included a number of new diagnostics to better understand and diagnose the impacts of the differing cumulus schemes on the resulting forecasts. Diagnostics such as area-averaged precipitation accumulation, precipitation partition, cloud fraction at low-, middle-, and high-levels, water budget components, and long- and short-wave radiation at the surface and top of atmosphere over specified regions (e.g., whole globe and Amazon) were calculated and plotted for the 10 June 2016 initialization. Comparisons were made both against CMORPH TMPA version 7 (Huffman et al. 2010). Using the case study approach, these additional diagnostics helped tie results from the SCM to the aggregated verification results, as well as provide further insight to the configurations behavior.

Verification

Objective model verification statistics were generated using the Model Evaluation Tools (MET) version 5.2; (http://www.dtcenter.org/met/users/docs/overview.php), and verification results plotted with the METViewer. Due to GFS-GF being initialized with operational data that uses the scale-aware SAS cumulus scheme for the first initialization in the test period, the first day of the experiment was discarded from the aggregated verification results. In the results shown later in the document, the time period is from 2 June - 15 June 2016.

For point-based verification, post-processed model output for surface and upper-air variables were compared to observations (METARs and RAwinsonde OBservations - RAOBs) using the MET point-stat tool. The 0.25° model output was regridded to G218, a 12-km Lambert Conformal grid covering the Contiguous United States (CONUS) and evaluated against NAM Data Assimilation System (NDAS) quality controlled BUFR (PrepBUFR) files for the surface verification. For upper-air verification, the 0.25° model output was regridded to both the G218 and G3 (a global 1.0° latitude-longitude domain) and evaluated using NCEP's Global Data Assimilation System (GDAS) PrepBUFR files as the observational dataset. Bias (or mean error) and RMSE were computed separately for each variable at the surface and upper-air levels. For the surface variables, statistics were aggregated over the CONUS domain, CONUS-WEST, CONUS-EAST, and 14 additional sub-regions. For brevity of the report, the focus is on CONUS results; the sub-region results are available via the project webpage. Upper-air statistics over global, CONUS, Northern

Hemisphere (NH; 20^{o} – 80^{o} N), Southern Hemisphere (SH; 20^{o} – 80^{o} S), Tropics (TROP; 20^{o} S – 20^{o} N), and Amazon domains for G3 are all available on the report web site.

Precipitation verification was performed over the entire globe. For the CONUS domain, a grid-to-grid comparison was made using the QPE from the Climatology-Calibrated Precipitation Analysis (CCPA) dataset, which has a resolution of ~4.8 km. Both the CCPA QPE analyses and the 0.25^{o} post-processed model output was interpolated to G218. For the global evaluation, CMORPH precipitation analyses (60^{o} N-60^{o} S) was used due to its high spatial (8 km at the equator, ~0.07) and temporal resolution. Both the CMORPH analyses and the 0.25^{o} post-processed model output was interpolated to G3 and compared over the NH (20^{o} – 60^{o} N), SH (20^{o} – 60^{o} S), Tropics (20^{o} S – 20^{o} N), and a region over the Amazon. Precipitation verification focused on 24-h accumulation period (valid from 12 UTC to 12 UTC) as well at the 6-h accumulations using the MET grid-stat tool. Traditional verification metrics computed for both CONUS and global regions include the frequency bias and the ETS.

Anomaly correlation, a measure of the ability of an NWP model to forecast synoptic-scale weather patterns (e.g., high-pressure ridges and low pressure troughs), as well as the location of frontal and storm systems, is a well-accepted verification metric used among operational centers and the research community. To compute the AC, the mean climatology was removed from the forecast and observations so that the strength of the linear association between the forecast and observed anomalies can be evaluated. The climatology files used for this test are the same 1.0^{o} GRIB1 files that are currently being used by NCEP. In order to pair the gridded forecast and analyses files with the climatology, the 0.25^{o} post-processed global forecasts were read into MET's grid-stat tool and then re-gridded to a 1.0^{o} grid before performing the AC calculation. Following accepted evaluation methods, the analysis files used in the evaluation were from the individual, respective model runs (i.e., GFS-SAS forecasts were compared against the GFS-SAS analysis files).

Since both configurations were run over an identical set of forecasts, the pairwise difference methodology is applied, when appropriate. With this methodology, differences between the verification statistics were computed by subtracting GFS-SAS from GFS-GF. The CIs on the pairwise differences between statistics for the two configurations objectively determine whether the differences are SS. For surface and upper-air, both the individual and pairwise verification statistics are accompanied by CIs computed from standard error estimates. The CIs were computed on the median values of the aggregated results for the surface and upper-air statistics using parametric tests. For the precipitation statistics, the bias-corrected and accelerated bootstrap method (using 1500 replicates) was used. The CIs on the pairwise differences between statistics for two configurations will assist in determining whether the differences are statistically significant. All CIs were computed at the 95% level.

With a copious amount of verification results being produced from this test, a "scorecard" is a straight-forward way to identify patterns in the difference of performance between two configurations, including level of significance, for specified metrics, variables, levels, regions, and times. This report includes verification results from the DTC's newly developed scorecard capability within METViewer. Note that this initial capability computes the means of the differences for continuous statistics, while the

surface and upper-air verification discussed above calculates the medians of the differences. As the scorecard capability matures, more user options will be included.

Verification of TCs was restricted to occurrence of genesis because the sample size was too small for verification of track and intensity of observed storms. Results from GFS-SAS, GFS-GF, and GFS-GFcold were compared against observed genesis as reported in the Best Track (BT). TC genesis locations were defined as the first time that NHC or JTWC designated the storm as a tropical depression (TD) or tropical storm (TS). The TC genesis verification criteria used in Halperin et al. (2016) was utilized with a slight modification. Here, a successful genesis forecast (i.e., "hit") was defined when observed genesis occurred within 240 h of the model initialization time and when the forecast genesis was within 5° of latitude and longitude of the BT location at the corresponding time. This spatial radius was used to account for uncertainties in the BT dataset, considered to be approximately 5°. For model genesis forecasts with valid times prior to the BT genesis time, Combined Automated Response to Query (CARQ) entries in the Automated Tropical Cyclone Forecasting (ATCF) system a-deck files were used to verify the forecast TC location. Genesis forecasts that were not observed, or that were present in the CARQ but not in the BT, were classified as false alarms (FA).

Initial Conditions, Forecast Periods and Length

Initial conditions for the global model were generated by running GDAS, which employs an 80-member T254 EnKF on a six-hourly cycle. GDAS was run separately for each model configuration (GFS-SAS and GFS-GF), so that initial conditions were consistent with the physics being tested.

Due to the addition of cycled DA, coupled with computational resource constraints, this test covered 15 forecasts within June 2016 (June 1 – June 15). While cycled DA was run every six hours, forecasts were launched once daily at 00 UTC and run out to ten days with output every six hours.

Scripts and Automation

Automation of tasks for this test were done using the Rocoto Workflow Management System, as currently employed in parallel tests at EMC. The scripts and automation system utilized the Rocoto-based gfs_workflow_v3.0.0 from EMC (https://svnemc.ncep.noaa.gov/projects/gfs/branches/gfs_workflow.v3.0.0, revision 91963), with additional verification and diagnostics tasks added by GMTB.

The xml file used to describe the tasks and their interdependencies was based on a Python-based xml generator developed by EMC. The created xml contains a variety of tasks, including setting up environment variables, creating the initial conditions by running GDAS, running the forecast model, post-processing, tracking tropical cyclones, and detecting tropical cyclogenesis. This preliminary xml was then augmented by GMTB with additional tasks to stage datasets, create forecast graphics, run forecast verification, archive results, and purge the disk. The various tasks (each of which evokes a number of executables) are submitted to the batch system incrementally as dependencies are met within the workflow. The modified xml as well as all other associated codes for running the global workflow portion

of the test resides in NOAA VLab under the "gmtb-tierIII" project name. Within this Git project, the specific code for running this test can be found in the master branch.

Archives

Output data files from multiple stages of the global workflow system were archived in the NOAA High-Performance Storage System (HPSS). The files are being saved to: /2year/BMC/gmtb/[configuration]/[YYYYMMDDHH], where configuration is sasda (GFS-SAS runs), gfda (GFS-GF runs), or gfff (GFS-GFcold runs) and YYYYMMDDHH is the initialization time in year-month-day-hour format. Archives include (file naming convention in parentheses):

- Configuration files and namelists specific to each forecast cycle (contained within [YYYYMMDDHH]_[configuration].tar)
- EnKF observation files from GDAS ([YYYYMMDDHH]gdas.enkf.obs.tar)
- Analysis files from GDAS for 00 UTC initializations ([YYYYMMDDHH]gdas.tar)
- Forecast files from GSM (analysis and forecasts at 6-hour increments; [YYYYMMDDHH]_[configuration]_bin1.tar and [YYYYMMDDHH]_[configuration]_bin2.tar)
- 0.25^{o} GRIB2 forecast files from *unipost* (analysis and forecasts at 6-hour increments containing full output [YYYYMMDDHH]_[configuration].tar; subset of output [YYYYMMDDHH]grib_subset.tar)
- Graphics from Python plotting suite ([YYYYMMDDHH]_fig.tar) and diagnostic routines
- Output from MET ([YYYYMMDDHH]_vx.tar])
- Logs from the individual tasks within the workflow ([YYYYMMDDHH]logs.tar)

Key Findings from Single Column Model (SCM_KF)

SCM_KF1. For the strongly forced maritime case, the GFS-GF suite produces weaker convective tendencies and convective transport than GFS-SAS. This alters the relationship among the physics schemes within the suite, leading to the explicit microphysics scheme in GFS-GF to show a greater relative response to the forcing.

Figure 2 shows the mean profiles of tendencies of specific humidity due to the supplied forcing and the parameterizations averaged over the deep convective period. Figure 3 shows the same for the temperature tendencies with the addition of curves due to longwave and shortwave radiation. While there are no observed tendencies to compare with, these plots are useful for interpreting how the two suites respond to the same inputs and provide value for interpreting quantities that do have observational analogs. It is clear that the convective tendencies produced in the GFS-GF suite are generally weaker than those from the GFS-SAS suite. Less column-wise drying and heating due to the convection shifts the response by the rest of the physics suite, most notably in the microphysics scheme. The microphysics scheme in the GFS-GF suite "compensates" by increasing its role -- producing a larger share of the total condensate and precipitation.

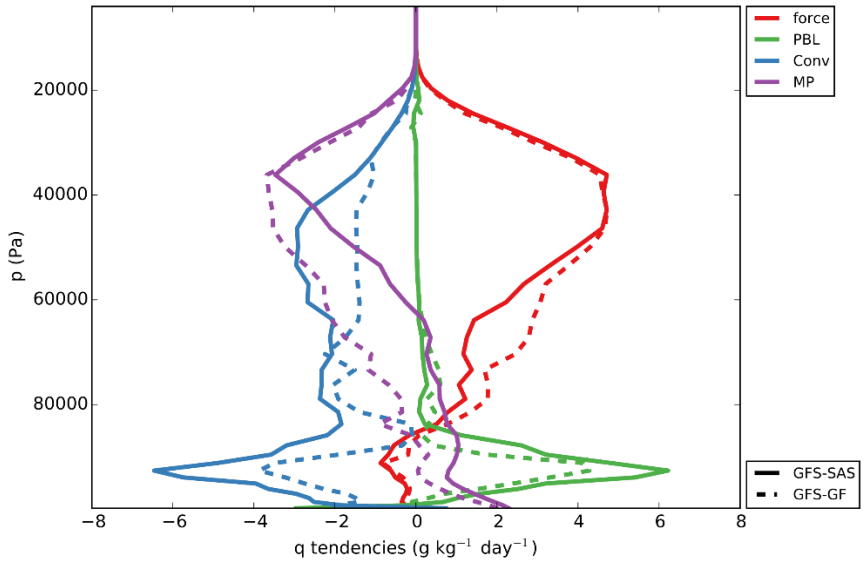


Figure 2. Mean profiles of specific humidity tendencies (g kg^{-1} day^{-1}) for the active phase of the TWP-ICE case. Colors denote forcing (red), PBL scheme (green), convective schemes (deep + shallow, blue), and microphysics scheme (purple). Line types denote the physics suite: GFS-SAS (solid) and GFS-GF (dashed).

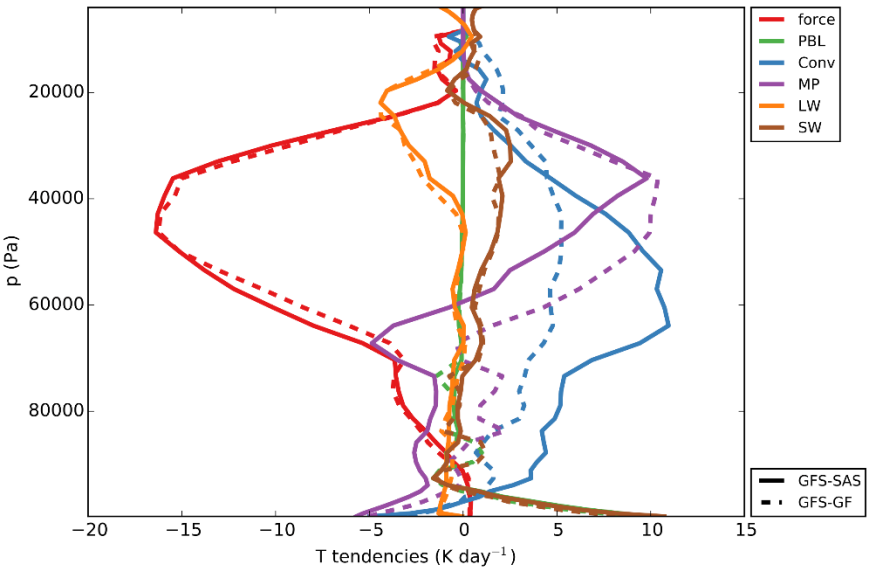


Figure 3. Same as Fig. 2 except for temperature tendencies (K day^{-1}). Tendencies due to longwave and shortwave radiation are in orange and brown, respectively.

Note that the forcing curves in these plots are not identical due to how forcing was applied for this case -- recall that the advective forcing terms are split into horizontal and vertical terms and only the former is explicitly prescribed. The vertical advective forcing depends on the modeled profiles of temperature and moisture, so any differences in the forcing profiles must be the result of differences in those profiles.

Figure 4 shows the mean convective mass fluxes averaged over the active convective period with red denoting updrafts, green denoting downdrafts, and blue denoting detrainment. The GFS-SAS suite produces a "bottom-heavy" profile of updraft mass flux, compared to the GFS-GF suite that produces reduced magnitudes below 400 hPa and a maximum in the middle troposphere. While both schemes generate bottom-heavy profiles of downdraft mass flux, the GF scheme's maximum value is approximately 10% of the SAS scheme. This provides evidence that the GF scheme is less active in vertical transport for this case, ultimately leading to the weaker convective tendencies shown in the previous two figures.

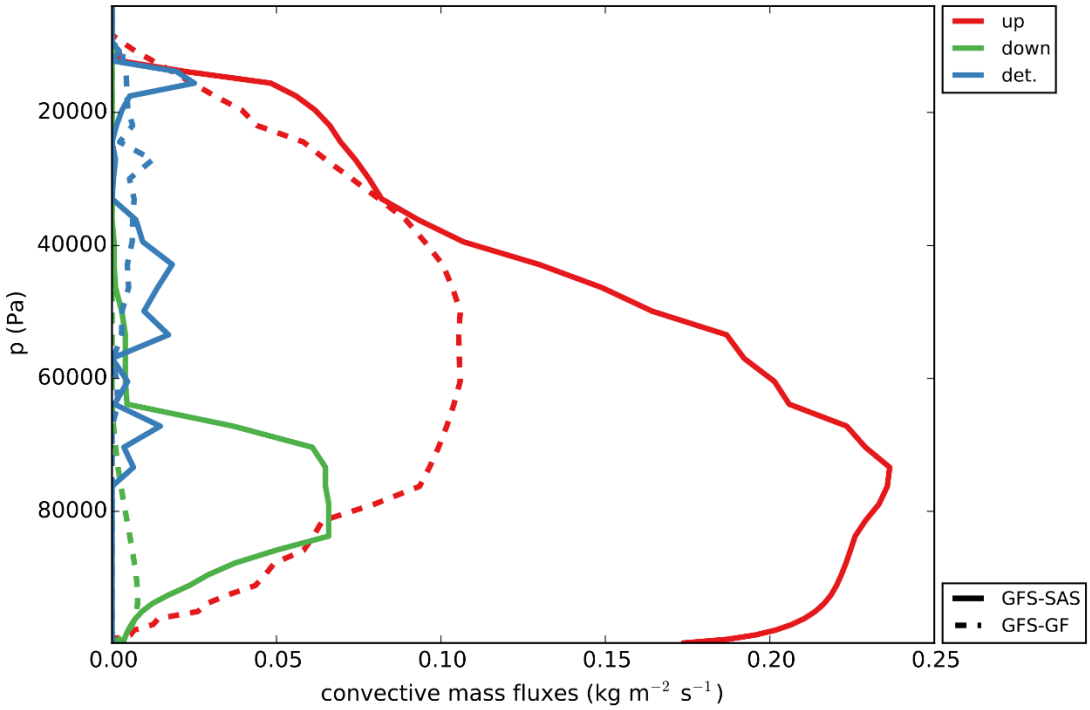


Figure 4. Same as Fig. 2 but for the convective mass fluxes (kg m-2 s-1). Colors denote mass flux type: updraft (red), downdraft (green) and detrainment (blue).

SCM_KF2. For the relatively weakly forced continental convection case, the convective tendencies produced by the GFS-GF suite were generally comparable to or greater than those produced by the GFS-SAS suite.

Figures 5-6 show the mean profiles of tendencies due to forcing and physics schemes for subperiod A of the continental ARM SGP case. First, notice how the moistening due to advective forcing was generally confined to below 600 hPa and less than 2 g kg^{-1} day^{-1}, and the temperature tendency due to advection exhibits weak heating in the PBL and modest cooling concentrated around 400 hPa and peaking around 5 K day^{-1}. Contrast this forcing with the TWP-ICE case that featured cooling aloft of more than 15 K day^{-1} coinciding with moistening of up to 4 g kg^{-1} day^{-1}. The response of the two suites to the weaker forcing differs considerably from the more strongly forced case. The GFS-GF suite produces stronger convective tendencies (or similar magnitude for subperiod B, not shown) for this case. The GFS-SAS suite produced weaker convective tendencies likely due to the fact that it was active and producing nonzero surface rainfall rates for a shorter percentage of the time. The suite with SAS tended to "cycle"

on and off more than the GF. Interestingly, the microphysical tendencies are of opposite sign for the strongly-forced maritime case compared to the weakly-forced continental case; the microphysics scheme is adding to the condensation of clouds in the maritime case, but is active in evaporating anvils (detrained condensate) in the continental case.

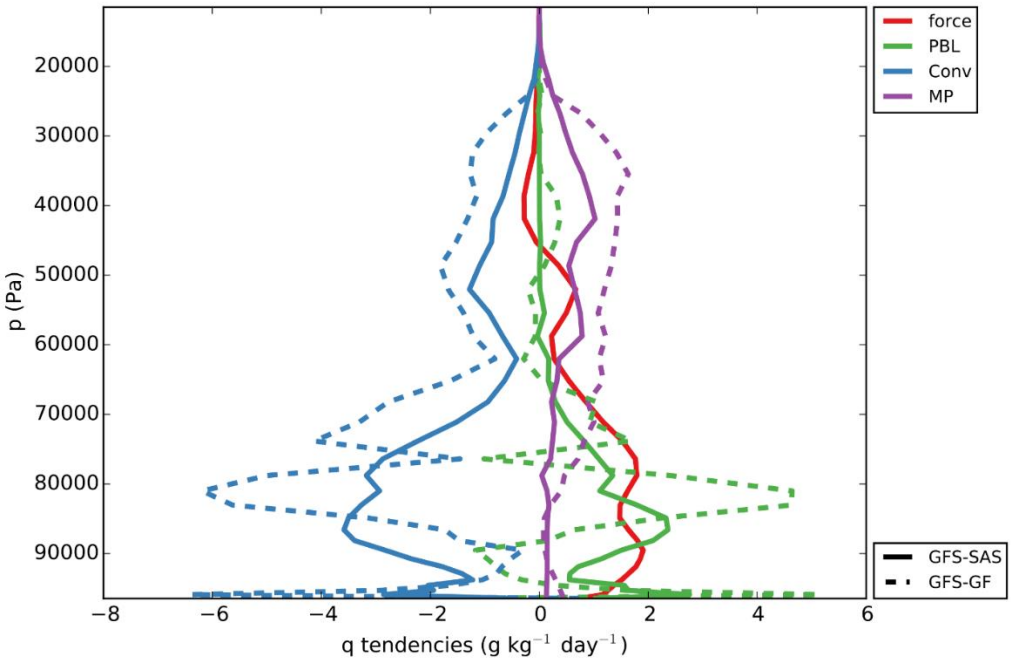


Figure 5. Same as Fig. 2 except for subperiod A of the continental ARM SGP case.

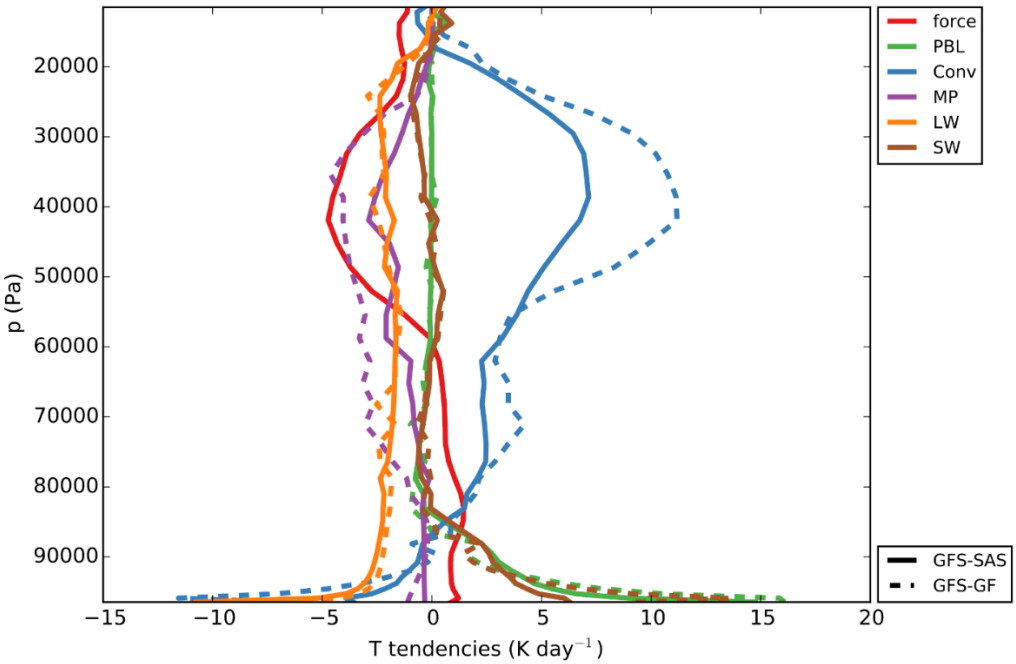


Figure 6. Same as Fig. 3 except for subperiod A of the continental ARM SGP case.

SCM_KF3. Use of the GFS-GF suite leads to higher moisture content in the boundary layer and generally produces a higher cloud fraction throughout the column, particularly in the lower-to-mid troposphere.

Figures 7-8 and Figs. 9-10 show mean profiles of specific humidity and cloud fraction, respectively, compared to observations for the active deep convective period of the TWP-ICE case and subperiod A of the ARM SGP case, respectively. In all cases and analysis periods, the GFS-GF suite produced between 0.5 and 1.5 g kg^{-1} higher values of water vapor specific humidity below about 700 hPa. For the TWP-ICE case and one subperiod of the ARM SGP case, this represented a reduction in the dry bias found in the GFS-SAS suite, while for the other subperiods of the ARM SGP case, this represented a flip from negative to positive water vapor bias. Higher levels of water vapor at these levels translated into higher values of cloud fraction and the associated reflected shortwave radiation. Taylor skill scores for water vapor and cloud fraction profiles were generally much improved through use of the GFS-GF suite.

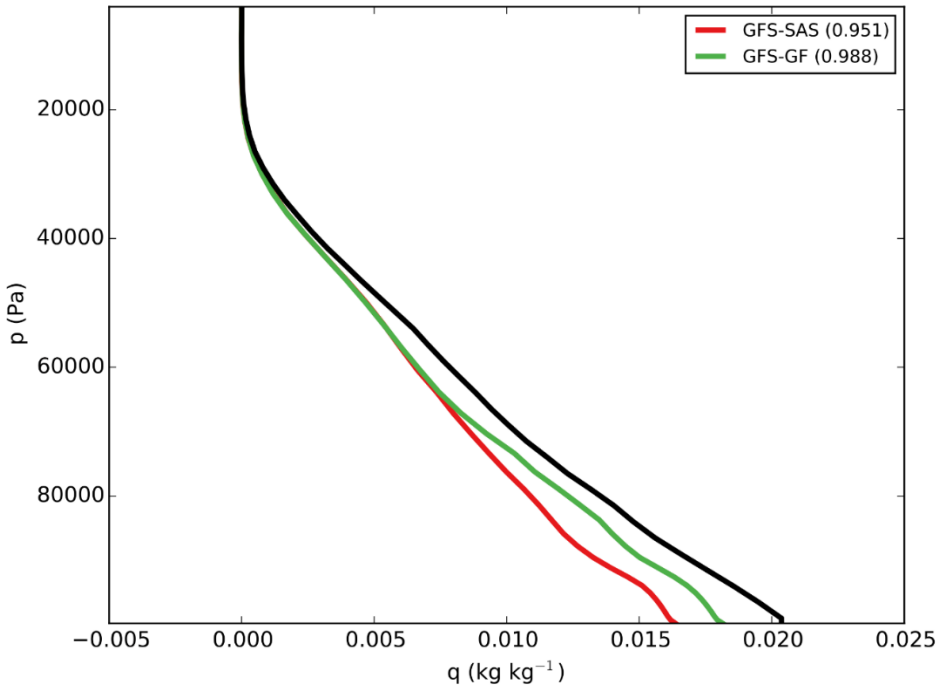


Figure 7. Mean profiles of specific humidity (kg kg^{-1}) averaged over the active phase of convection for the TWP-ICE case. Colors denote observations (black) or physics suite (GFS-SAS in red and GFS-GF in green). Skill scores are printed in the legend.

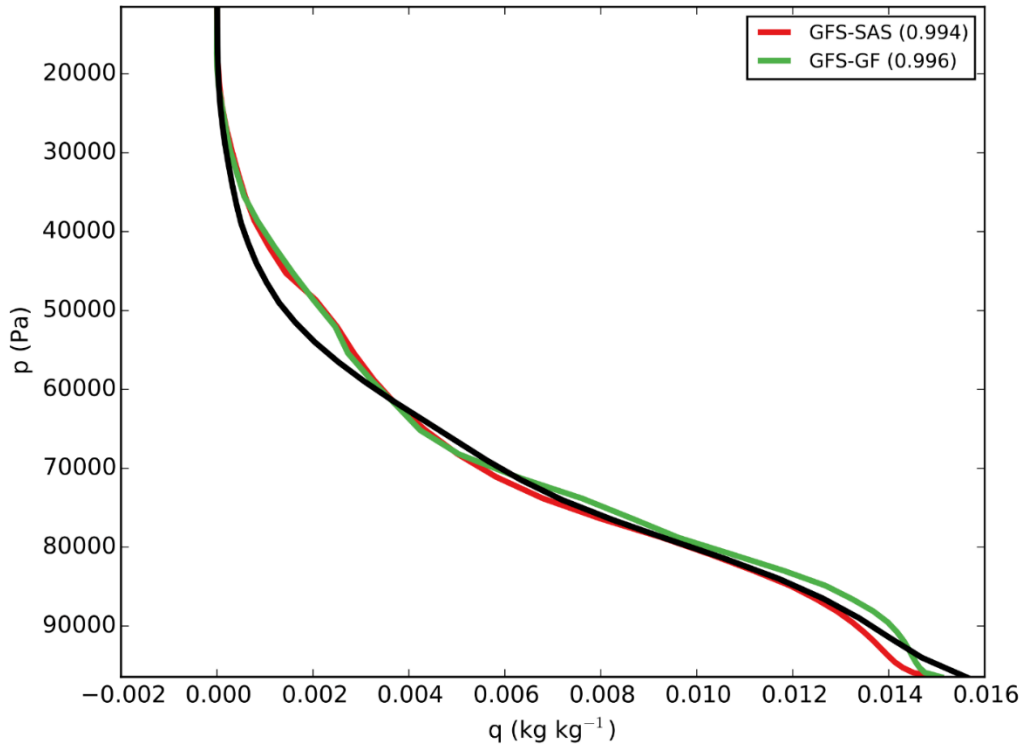


Figure 8. Mean profiles of specific humidity (kg kg^{-1}) averaged over the active phase of convection for the TWP-ICE case. Colors denote observations (black) or physics suite (GFS-SAS in red and GFS-GF in green). Skill scores are printed in the legend.

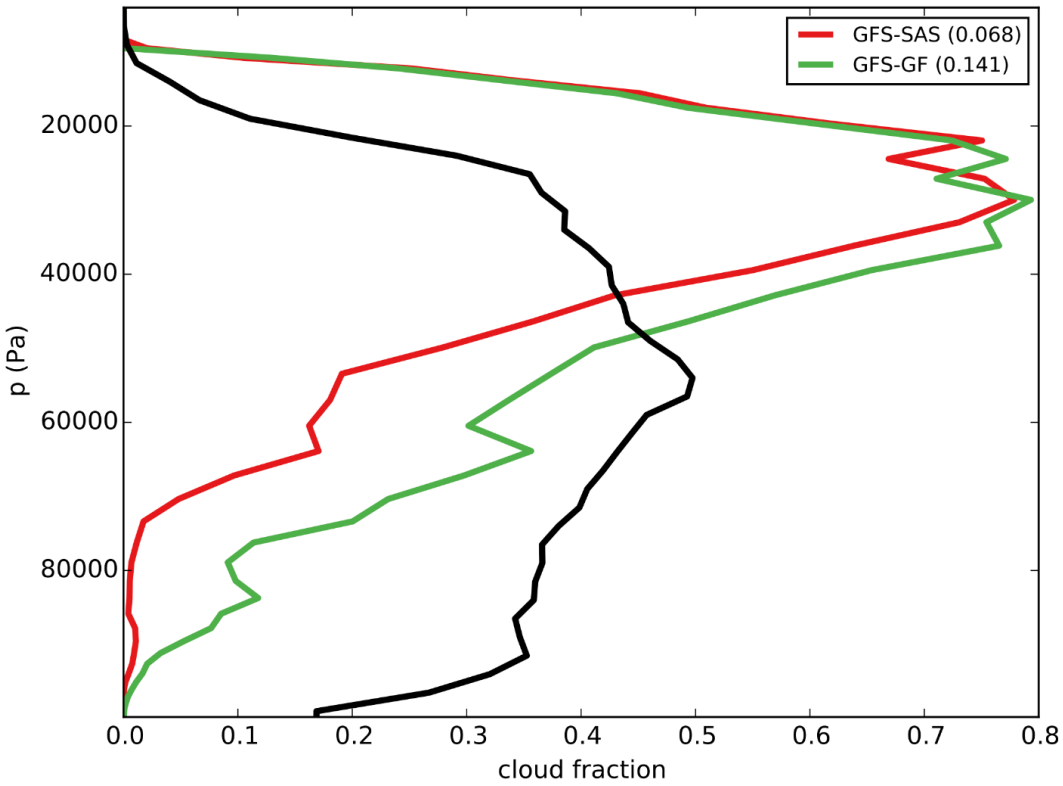


Figure 9. Same as Fig. 7 except for cloud fraction.

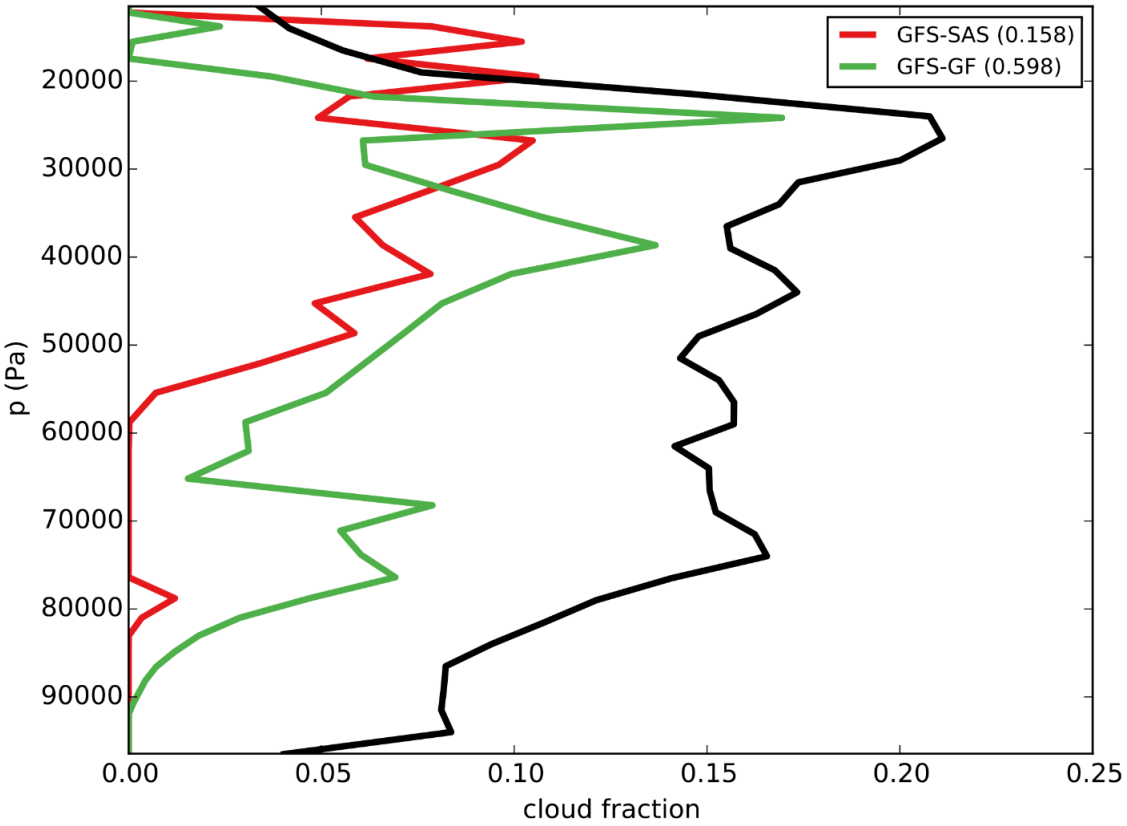


Figure 10. Same as Fig. 9 except for subperiod A of the ARM SGP case.

SCM_KF4. During the suppressed convection phase of the maritime convective case and two subperiods of the continental convective case, the GFS-GF suite alters the interaction with the PBL scheme, leading to the transport of PBL moisture higher in the column and occasionally spuriously large cloud fraction at the PBL top.

Figures 11-12 and Figs. 13-14 show mean profiles calculated from the suppressed convective period of the TWP-ICE case and subperiod B of the ARM SGP case. The profiles of the tendencies of water vapor show large changes in where the PBL transport "deposits" its moisture. The elevated moisture content at the top of the PBL leads to spurious cloud formation for these two analysis periods. The thicker PBL cloud relocates the position of maximum shortwave radiative cooling higher in the column (not shown). For the suppressed convection phase of the TWP-ICE case, the combination of cooling from the PBL scheme and reflected shortwave even creates a spurious shallow temperature inversion around 700 hPa (not shown).

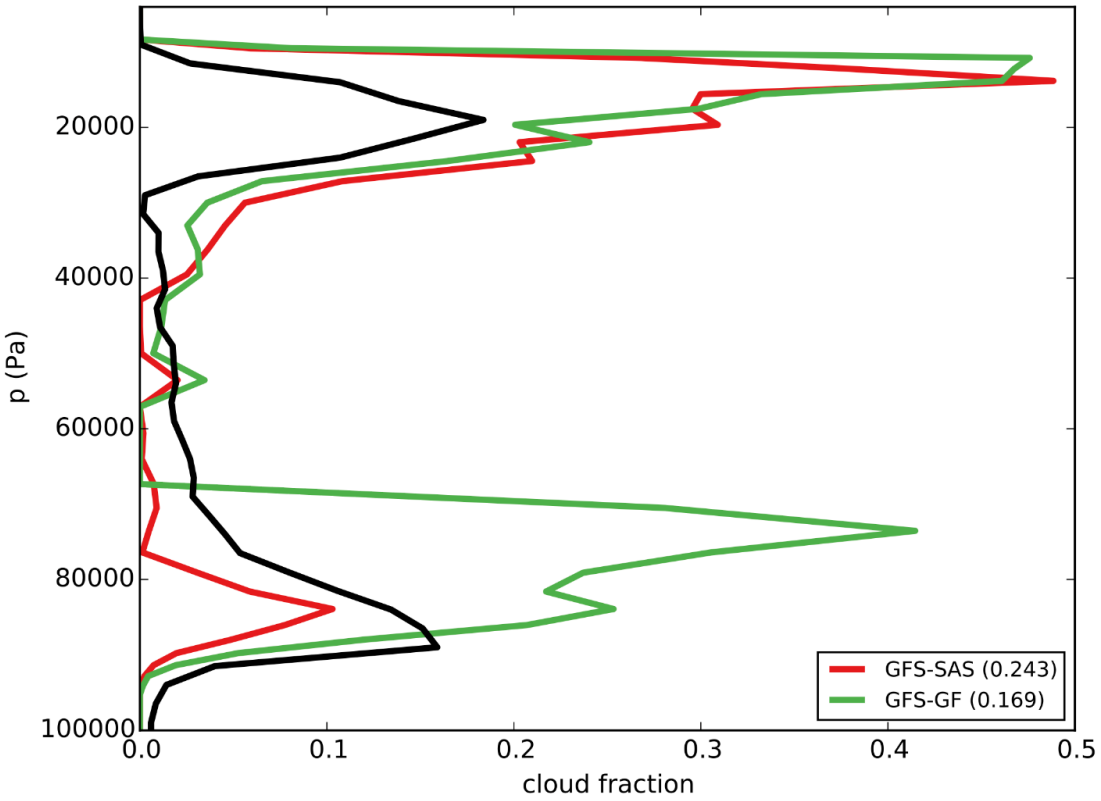


Figure 11. Mean profiles of cloud fraction averaged over the suppressed phase of convection for the TWP-ICE case. Colors denote observations (black) or physics suite (GFS-SAS in red and GFS-GF in green). Skill scores are printed in the legend.

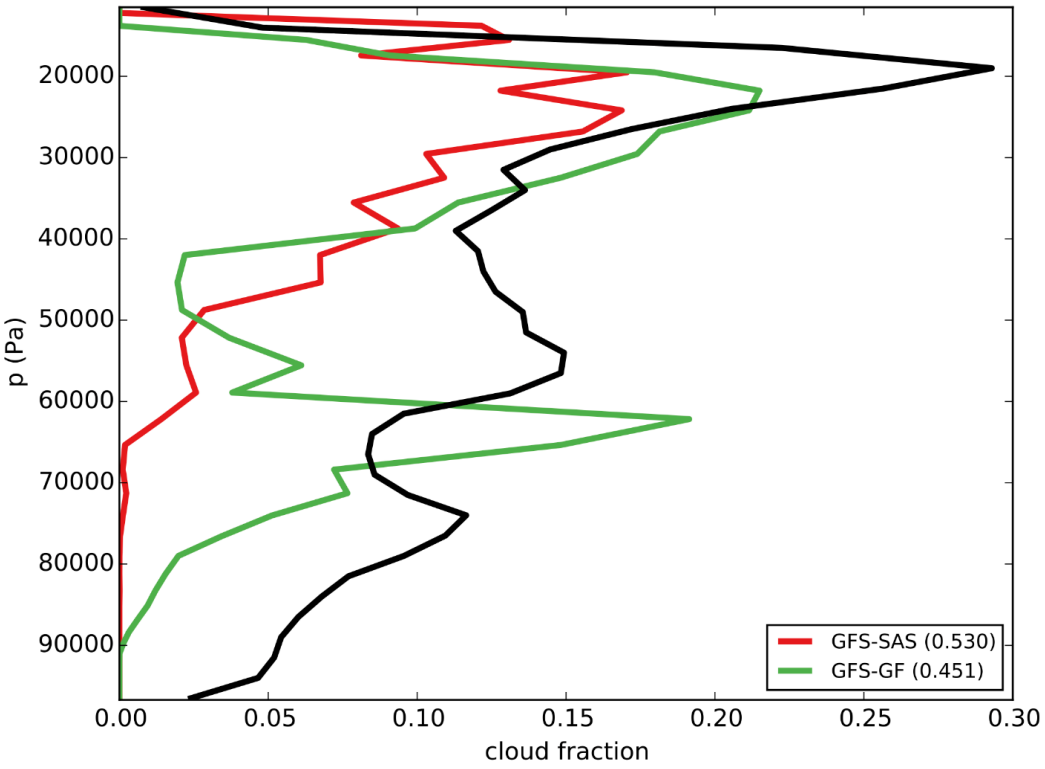


Figure 12. Same as Fig. 11 except for subperiod B of the ARM SGP case.

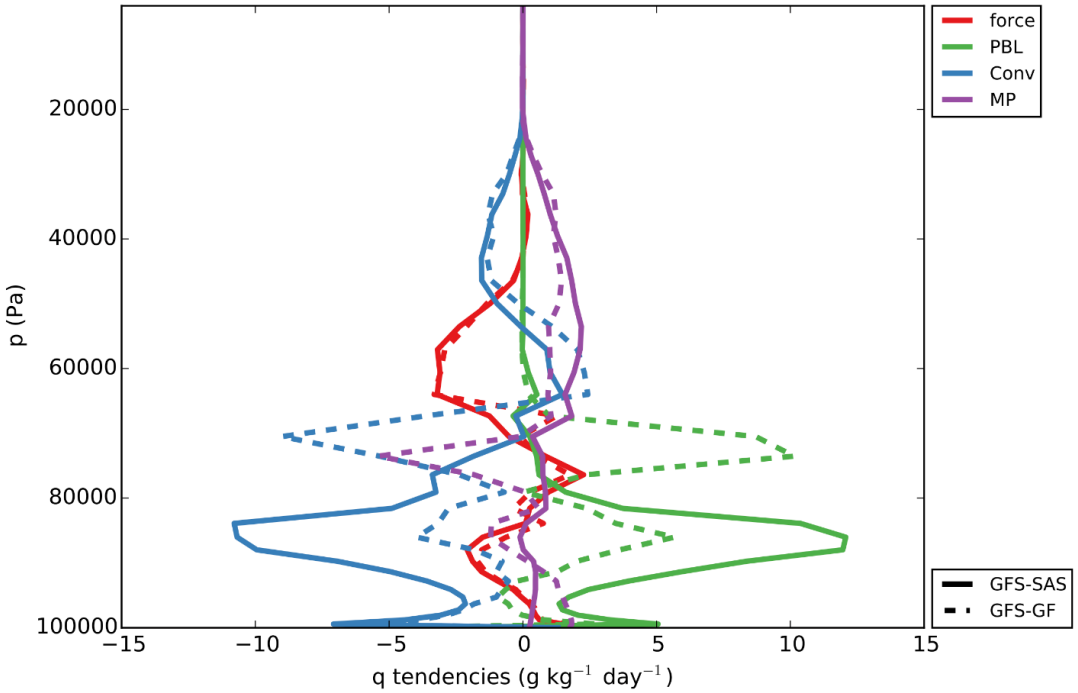


Figure 13. Mean profiles of specific humidity tendencies (g kg^{-1} day^{-1}) for the suppressed phase of the TWP-ICE case. Colors denote forcing (red), PBL scheme (green), convective schemes (deep + shallow, blue), and microphysics scheme (purple). Line types denote the physics suite: GFS-SAS (solid) and GFS-GF (dashed).

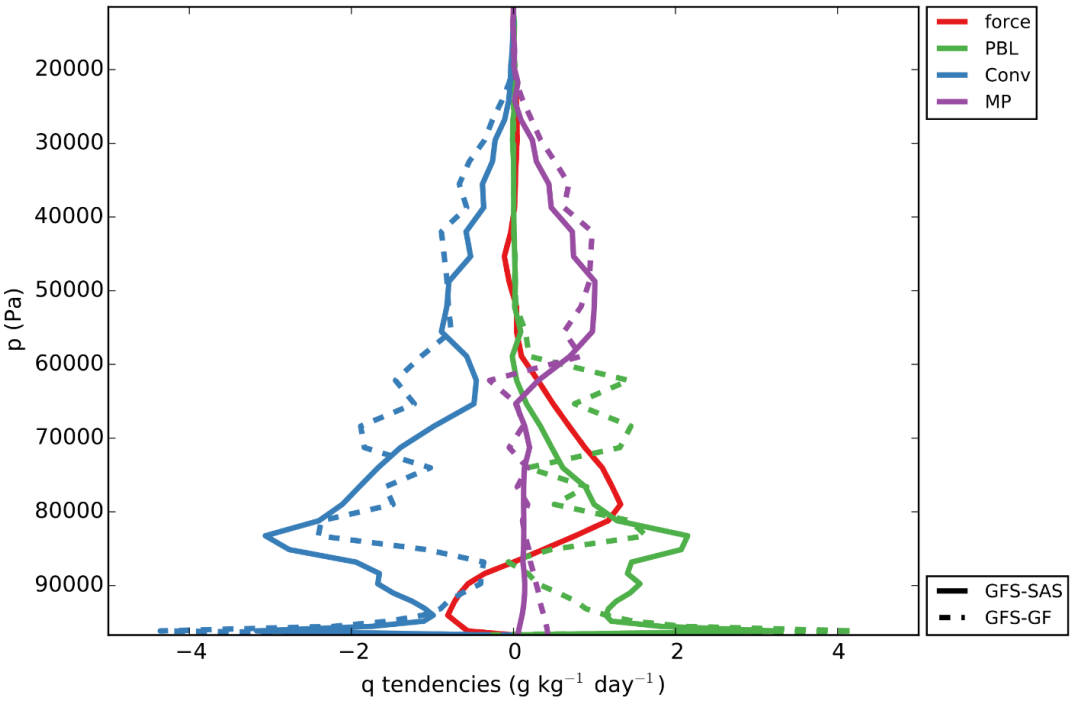


Figure 14. Same as Fig. 13 except for subperiod B of the ARM SGP case.

SCM_KF5. Although both suites produce approximately the same precipitation amounts for both cases, the GFS-GF suite produces a much lower convective precipitation ratio and lower temporal variability than the GFS-SAS suite.

Given abundant moisture, the total precipitation produced by a single-column model is largely a function of the imposed forcing. Details such as timing and partition into convective and explicit can provide insight into the physics. Figures 15-16 show the total surface precipitation rate time series during the active phase of the maritime case and subperiod A of the continental case (respectively) for both suites compared to observations. Note that although values for the SCM output are available for every timestep, the observed data points have a period of 3 hours; the SCM output is resampled to the observed period. Although total precipitation for both suites is within about 5% of each other, the GFS-SAS suite has a slightly higher Taylor skill score for the maritime case and a much higher score for this subperiod of the continental case, indicating higher temporal correlation, more similar variability, or both.

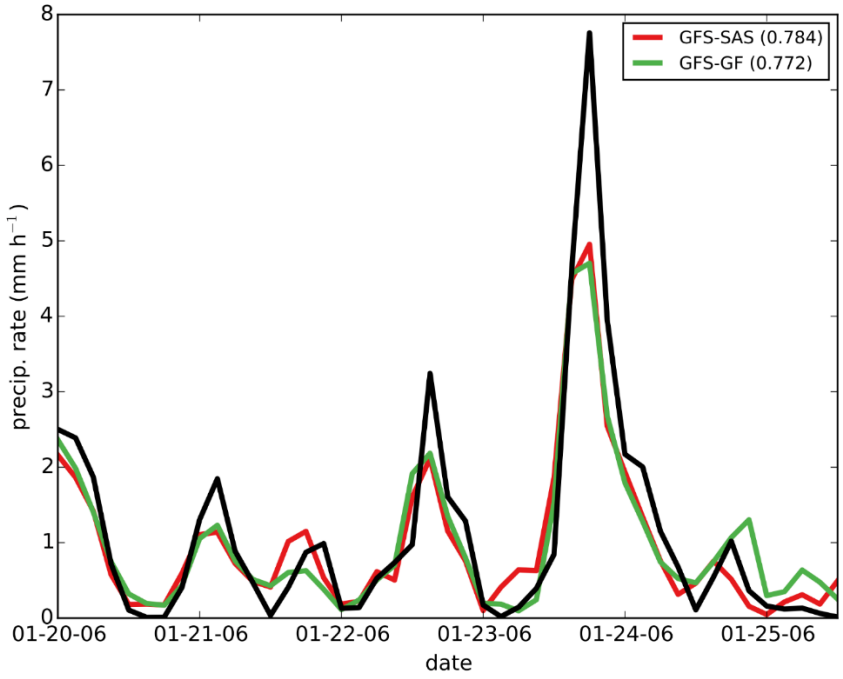


Figure 15. Time series of total surface precipitation rate (mm h^{-1}) for the active phase of the TWP-ICE simulation. Colors denote observations (black) or physics suite (GFS-SAS in red and GFS-GF in green). Skill scores are printed in the legend.

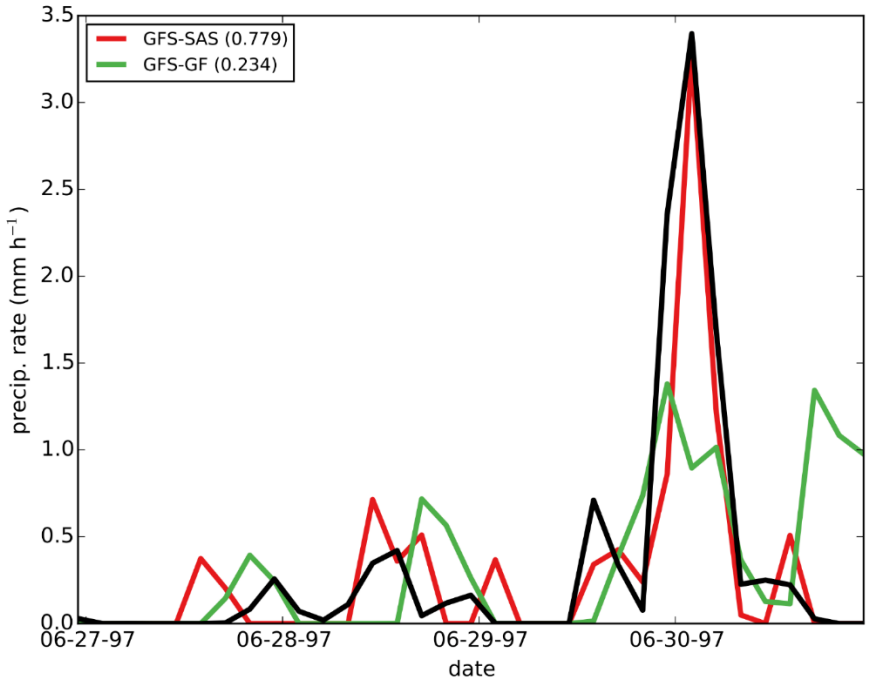


Figure 16. Same as Fig. 15 except for subperiod A of the ARM SGP case.

The biggest difference between the suites for the maritime case in terms of precipitation, however, is the partition between convective and explicit sources. Figures 17-18 show scatter plots of total precipitation rate versus the ratio between convective and total precipitation, with each point

representing the time average for one of the 100-member forcing ensemble for the deep convective period (Fig. 17) and shallow convective period (Fig. 18) of the maritime case. The total precipitation rate is a proxy for the strength of the applied forcing. We find that the convective ratio is always lower for the GFS-GF suite than for the GFS-SAS suite, and the GFS-GF exhibits a stronger relationship to the applied forcing. For the deep convective period, an apparently discrete "jump" in the ratio for the GF scheme exists where the convective ratio goes from a much lower value to a value that is much closer to the GFS-SAS suite. For the suppressed convection, only a weak relationship between forcing strength and convective ratio is evident for the GFS-SAS suite, but a nearly monotonically increasing relationship is exhibited for the GFS-GF suite.

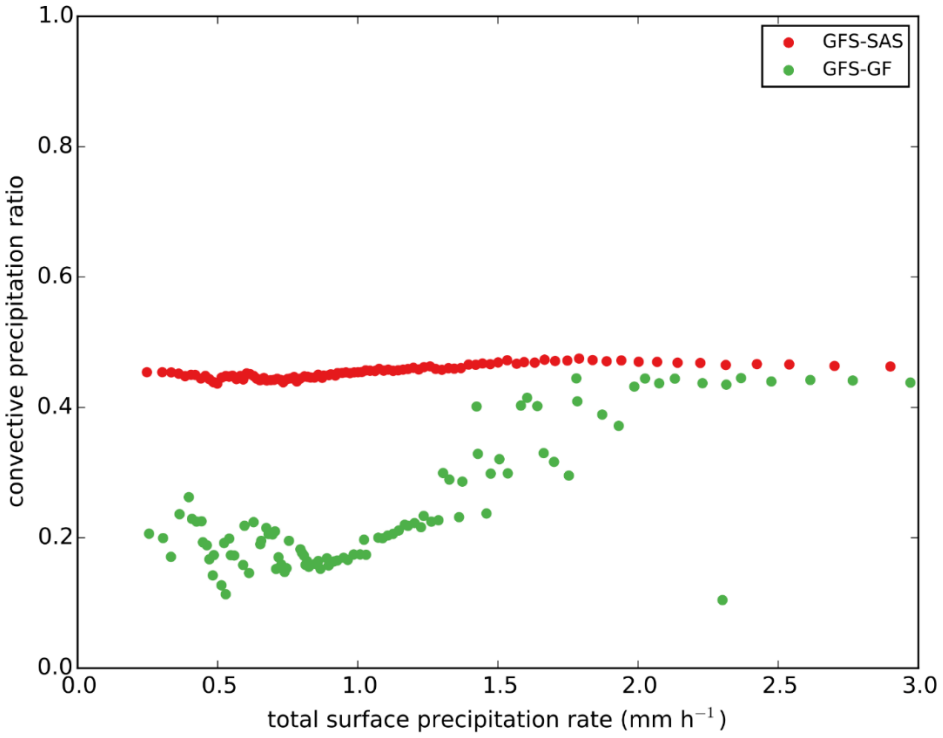


Figure 17. Scatter plot denoting mean values of total surface precipitation rate (mm h^{-1}) versus convective precipitation ratio over the active phase of convection for all forcing ensemble members for GFS-GF (green) and GFS-SAS (red).

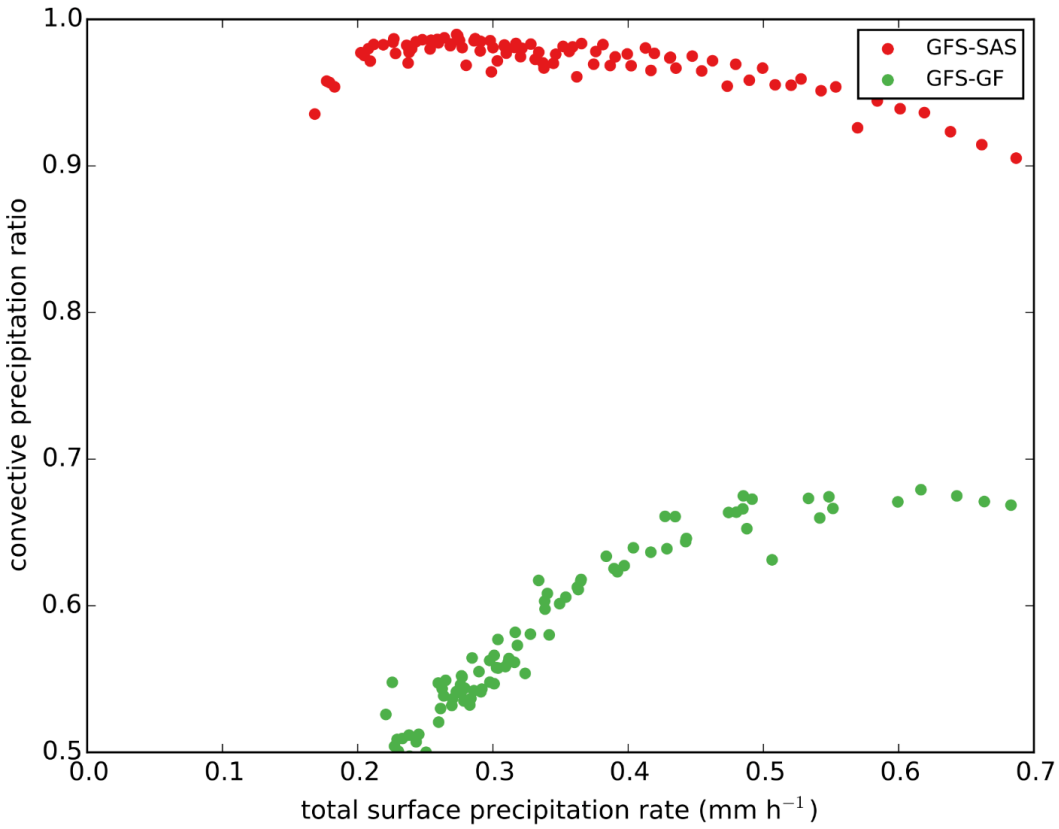


Figure 18. Same as Fig. 17 but for the suppressed phase of convection.

Figure 19 shows similar plots from the SCM intercomparison of Davies et al. (2013). The GFS physics suite from circa 2011 (labeled "NCEPG" in light blue) exhibits nearly identical behavior to the control GFS-SAS suite presented here. The GFS-GF suite appears to behave quite similarly to the GISS model found in that intercomparison. Further, Davies et al. (2013) identified two groupings of behavior, those with a relatively high convective ratio and those with a relatively low convective ratio. Switching from SAS to GF causes the GFS physics suite to switch from the "high" group to the "low" group. For completeness, this result holds for the continental case too. The convective precipitation ratio for the GFS-SAS suite for all subperiods of the continental case are between 0.97 and 1 whereas they are between 0.86 and 0.89 for the GFS-GF suite.

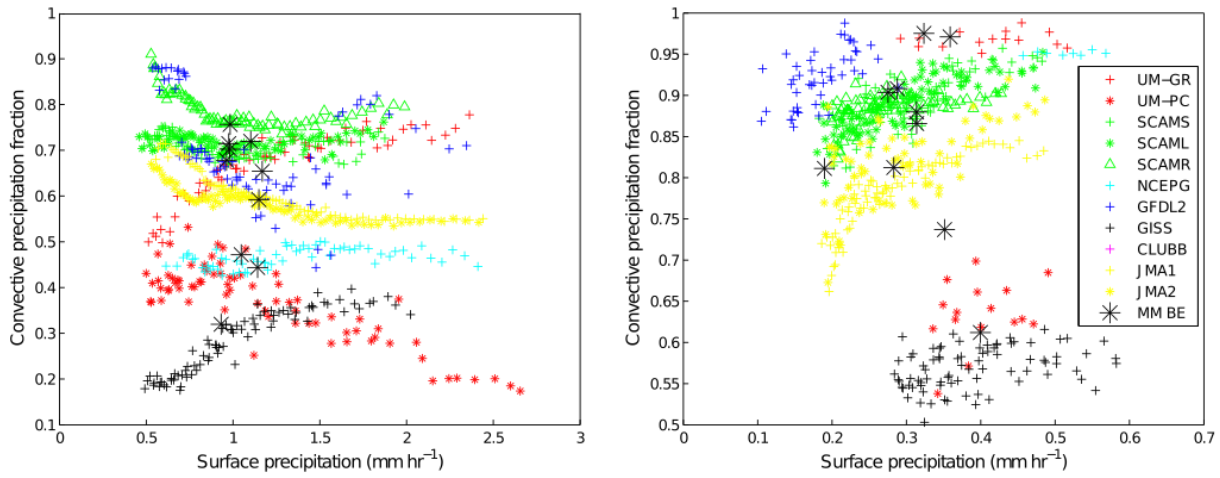


Figure 19. Fig. 7 from Davies et al. (2013). The format and axes are the same as for figures 17 and 18. In this plot, colors denote models that participated in a SCM intercomparison project, with the GFS from circa 2011 in light blue.

Figure 20 shows the same data as Fig. 16 for the continental case except that it shows surface precipitation *for each time step* compared to observations (not resampled to 3-hourly). This plot demonstrates a key difference between the two suites. The GFS-SAS suite shows considerably more temporal variability -- its convective scheme appears to be activated/deactivated much more frequently than the GFS-GF suite that tends to stay activated for the duration of a precipitation event. This is related to the lower convective precipitation ratio since the GFS-GF suite produces more grid-scale precipitation to temporally "smooth out" the surface precipitation time series.

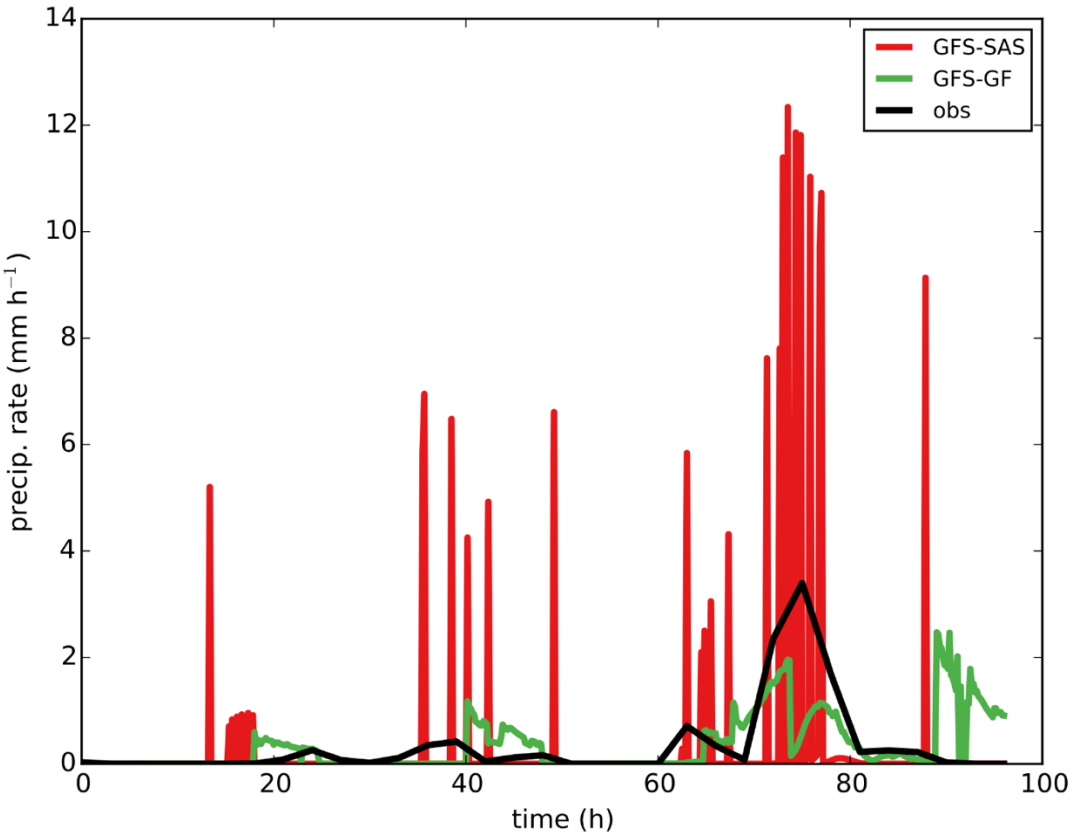


Figure 20. Same as Fig. 16 without resampling model data to the observation period. Data from each time step is displayed.

SCM_KF6. During the maritime deep convective period, the forcing ensemble elicits greater variability from the GFS-GF suite than the GFS-SAS suite.

Figures 21, 22, and 23 show mean profiles of specific humidity, total convective temperature tendency, and total convective moisture tendency, respectively, averaged over the deep convective phase of the maritime case for the forcing ensemble. These plots demonstrate that the GFS-GF suite is more sensitive than the GFS-SAS suite to the forcing. In fact, for these quantities, at most levels, the GFS-GF ensemble range contains the GFS-SAS ensemble range. For convective tendencies, the 25th percentile profiles for the GFS-SAS approximate the 50th percentile of the GFS-GF suite and the 50th percentile profile of the GFS-SAS suite approximates the 75th percentile profile of the GFS-GF suite for much of the depth of the atmosphere. However, for more extreme ends of the forcing ensemble range, the GFS-GF suite produces more extreme profiles than the GFS-SAS suite. The different response to changes in forcing likely has consequences for stochastic applications.

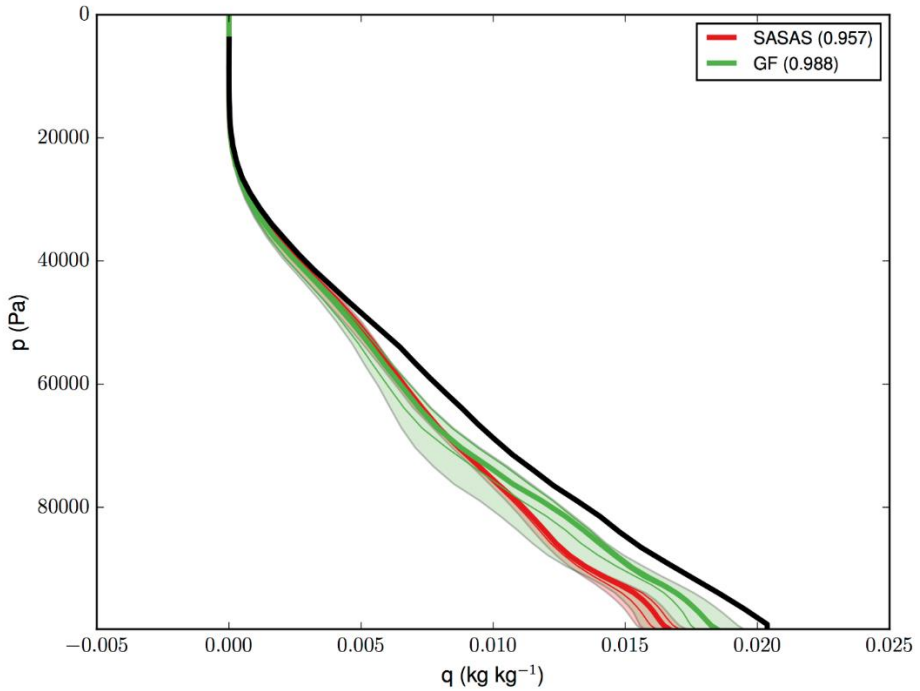


Figure 21. Mean profiles of specific humidity (kg kg^{-1}) for the active convective phase of the TWP-ICE case. Colors denote observations (black) or physics suite (GFS-SAS in red and GFS-GF in green). Shading encompasses the 10-90th percentiles of the forcing ensemble. Thin lines denote the 25th and 75th percentiles. Thick lines denote the 50th percentile.

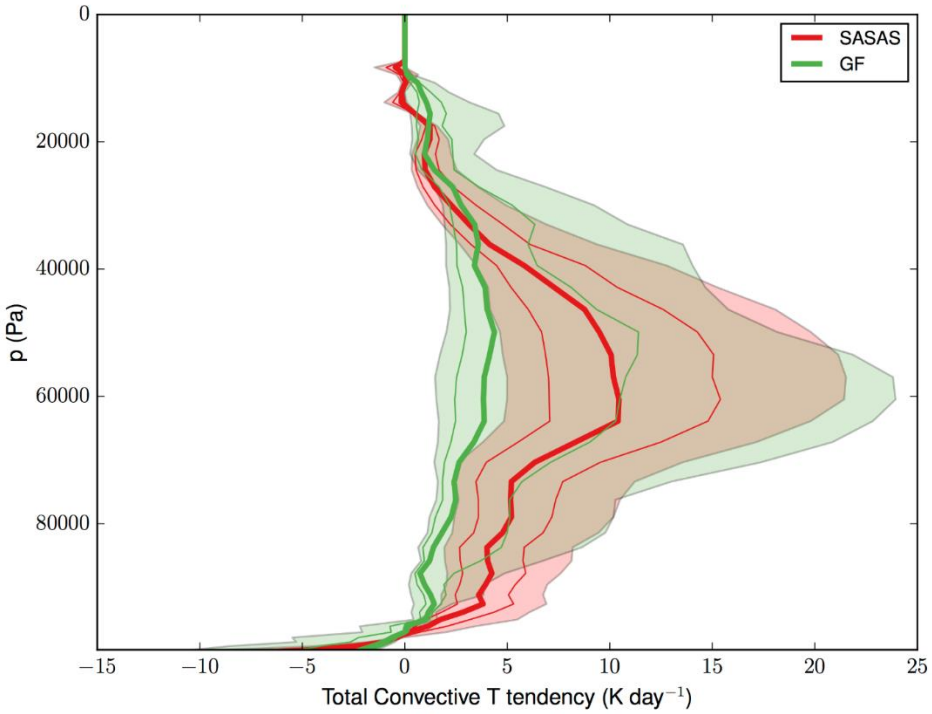


Figure 22. Same as Fig. 21 except for the total convective temperature tendency (K day^{-1}).

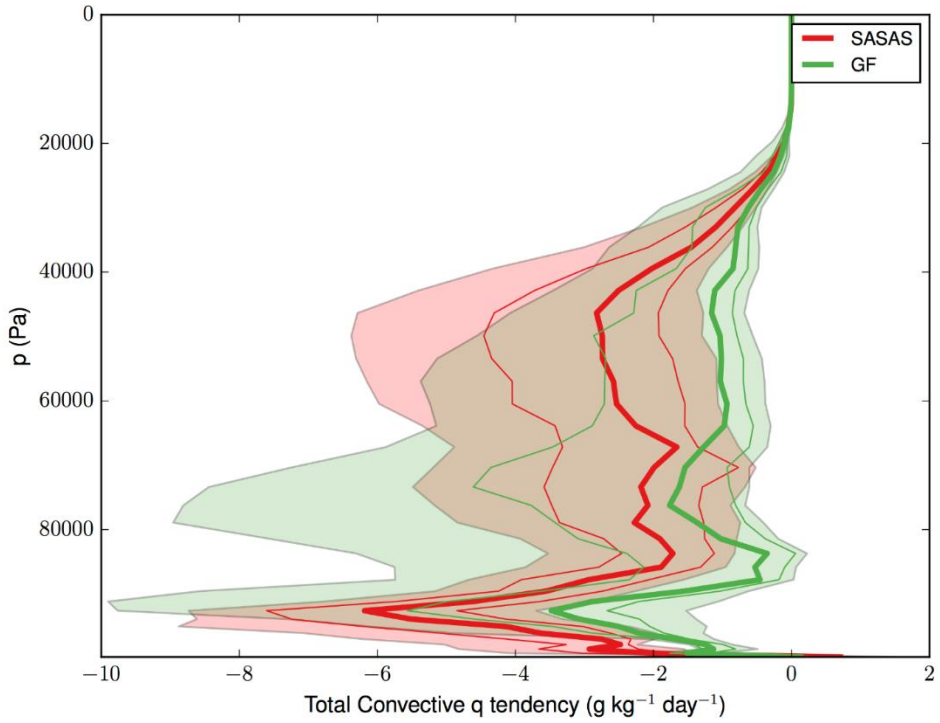


Figure 23. Same as Fig. 21 except for the total convective moisture tendency (g kg^{-1} day^{-1}).

Key Findings from Global Diagnostics (GD_KF) of Global Single Case

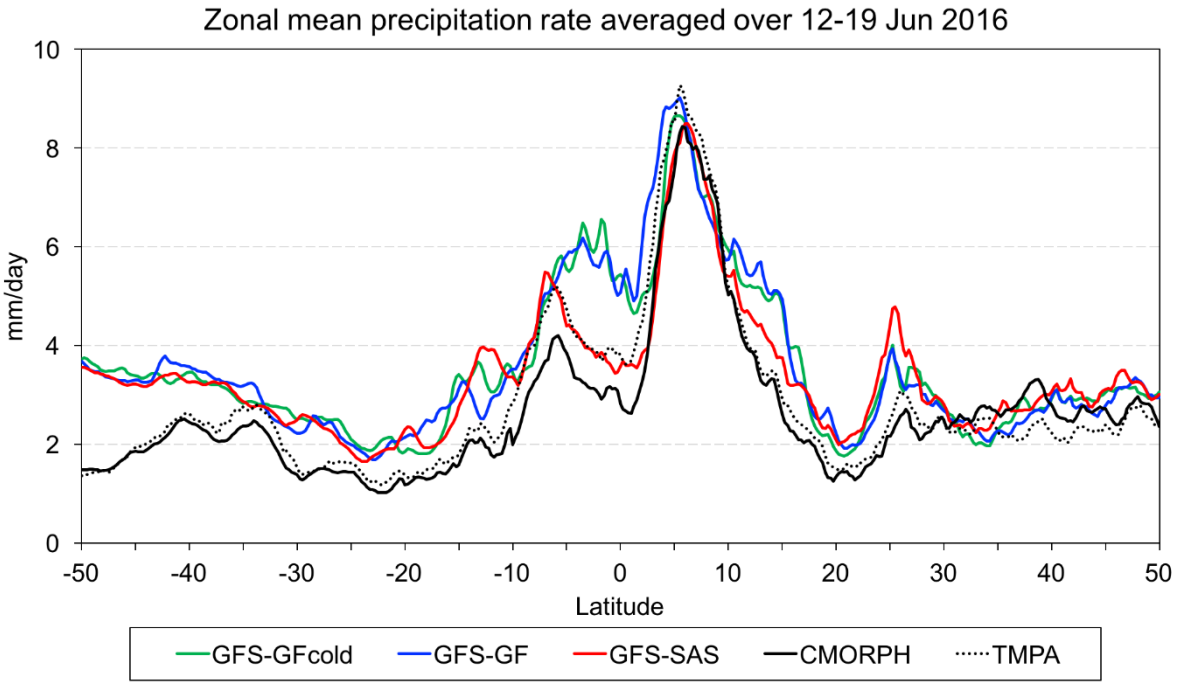
GD_KF1. GFS-GF produces extra precipitation in the tropics, especially between 5°S and 5°N.

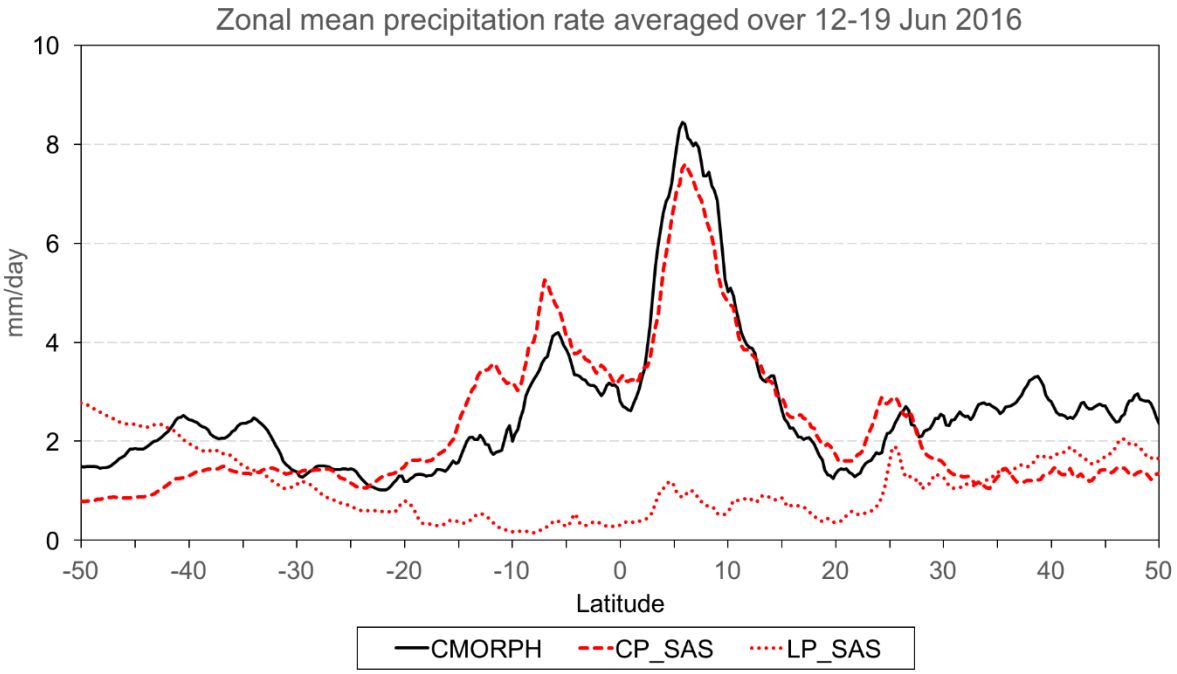
Compared to CMORPH, all configurations overestimate global precipitation, especially on the Southern Hemisphere and in the tropics between 5°S and 5°N. This problem is particularly pronounced for the configurations using GF (Fig. 24a), which fail to represent the minimum in precipitation along the equator between the ITCZ and the SPCZ on the Maritime Continent. It is interesting to note that TMPA has generally higher precipitation than CMORPH between 30°S and 30°N, considering that both of the two dataset have a spatial resolution of 0.25° and a temporal resolution of 3-hourly. Compared to TMPA, GFS-SAS represented reasonably the location and strength of SPCZ near 5°S.

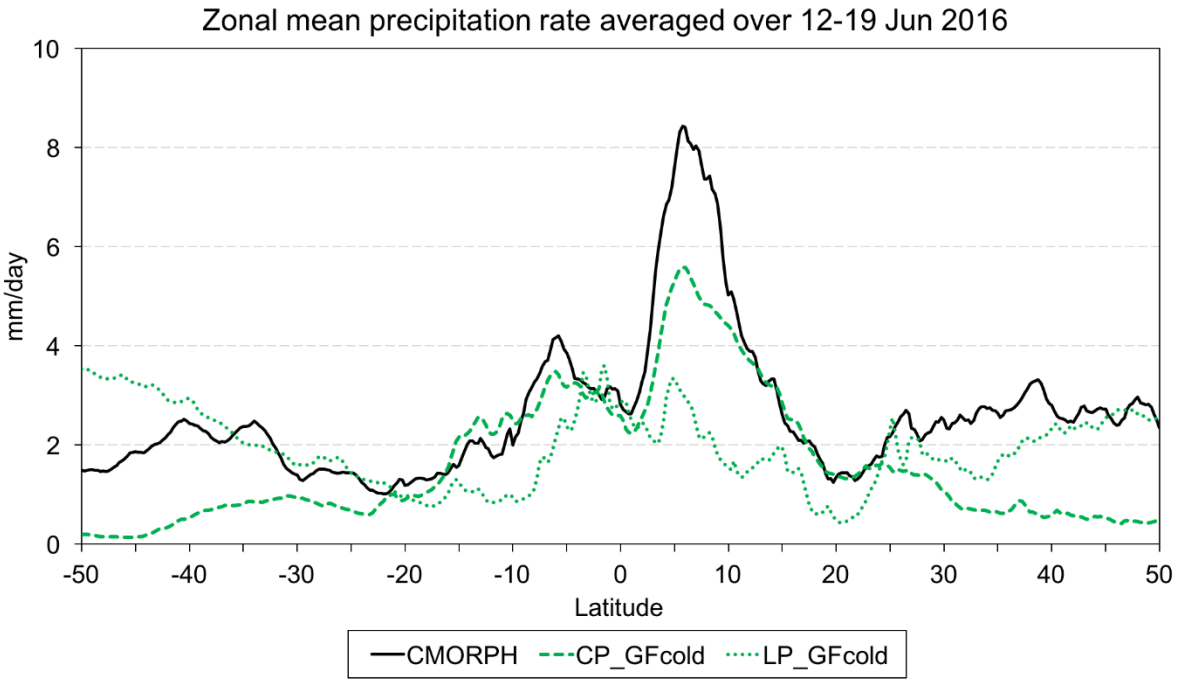
GD_KF2. Total precipitation, and its partition between convective and explicit components, is different between GFS-SAS and GFS-GF. Precipitation development occurs faster for the cycled runs. Comparing GFS-SAS with GFS-GF, the spin up time is shorter in GFS-SAS. Compared to CMORPH observations, model precipitation in GFS-GF is too light and frequent in rainfall intensity between 2-7 mm d^{-1}.

In the GFS physics suite, precipitation can be generated via grid-scale condensation (LSP) and cumulus parameterization (CP). Figures 24 and 25 show that the partitioning of precipitation between LSP and CP is strongly impacted by cumulus parameterization and region. A global view of the partitioning

(Fig. 25) indicates that the majority of precipitation amount in GFS-SAS is from the CP, but that in GFS-GF it is from the LSP.







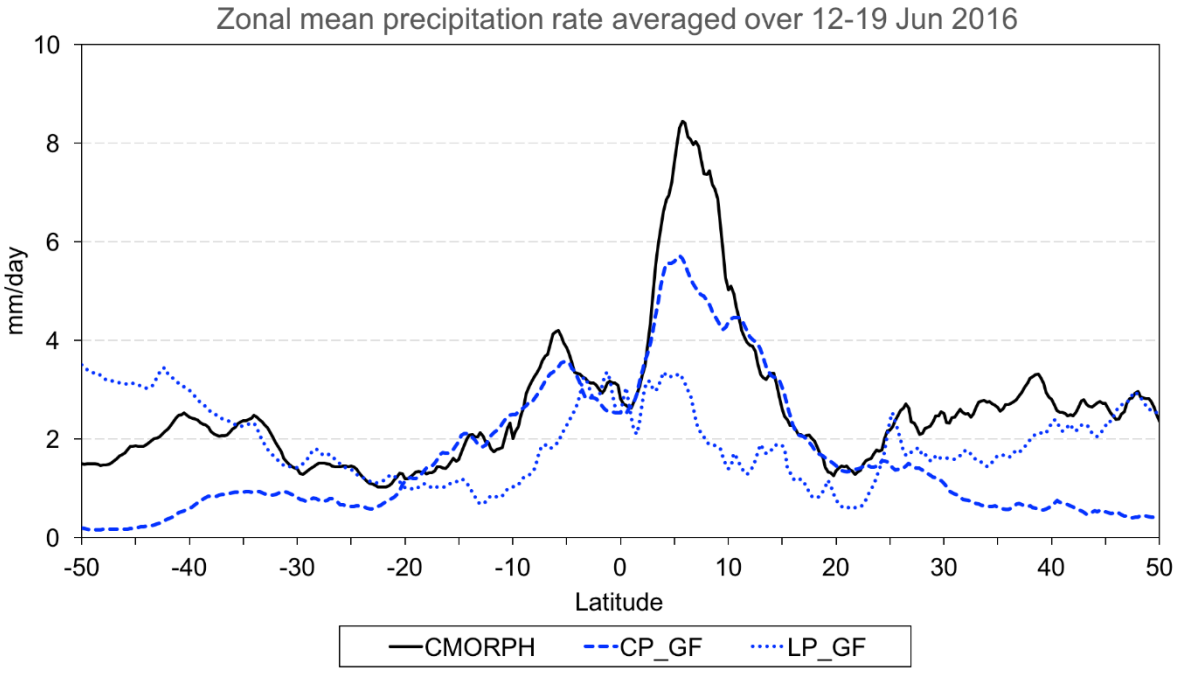


Figure 24. (a) Zonal mean of total precipitation rate (mm day^{-1}), CMORPH data and TMPA satellite data; and zonal mean of forecast convective (dashed line) and explicit (dotted line) precipitation rate (mm day^{-1}) in (b) GFS-SAS; (c) GFS-GFcold; (d) GFS-GF for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016. GFS-SAS is displayed in red, GFS-GFcold in green, GFS-GF in blue, and the CMORPH observations in black.

The zonal mean of precipitation rate in Fig. 24b-d shows that in the Tropics (20°S-20°N), precipitation in GFS-SAS is mostly CP, while LSP and CP components are closely matched in strength in GFS-GF. In the mid-latitudes, the LSP prevails in GFS-GF, while LSP and CP are matched in strength in GFS-SAS, especially in the Northern Hemisphere.

After initialization, precipitation develops faster for the cycled runs (Fig. 25a). This is particularly true for the convective precipitation (Fig. 25b) since a spin up in explicit precipitation is still noticed in GFS-GF (Fig. 25c) in the first 24 hours.

The joint assessment of model frequency and accumulation of precipitation provides more insight on the validity of model moist processes. Observed and forecast daily-accumulated precipitation rates between 50°S and 50°N (Fig. 25d) indicate that the frequency of occurrence of total precipitation in GFS-SAS is very similar to that of CMORPH below 7 mm d^{-1}. All configurations overestimate in the forecast frequency of precipitation in GFS-GF between 2-7 mm d^{-1} when compared to TMPA.

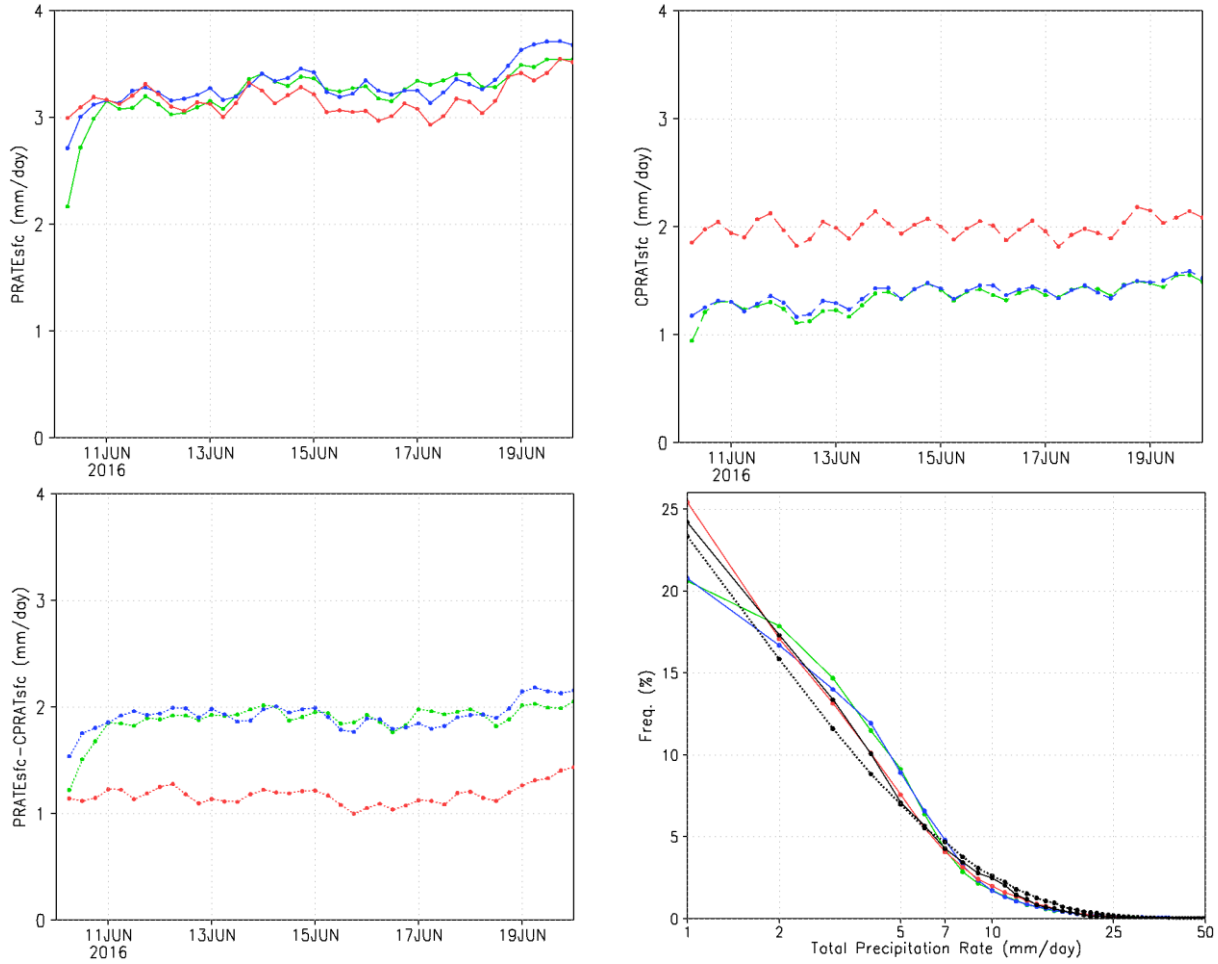
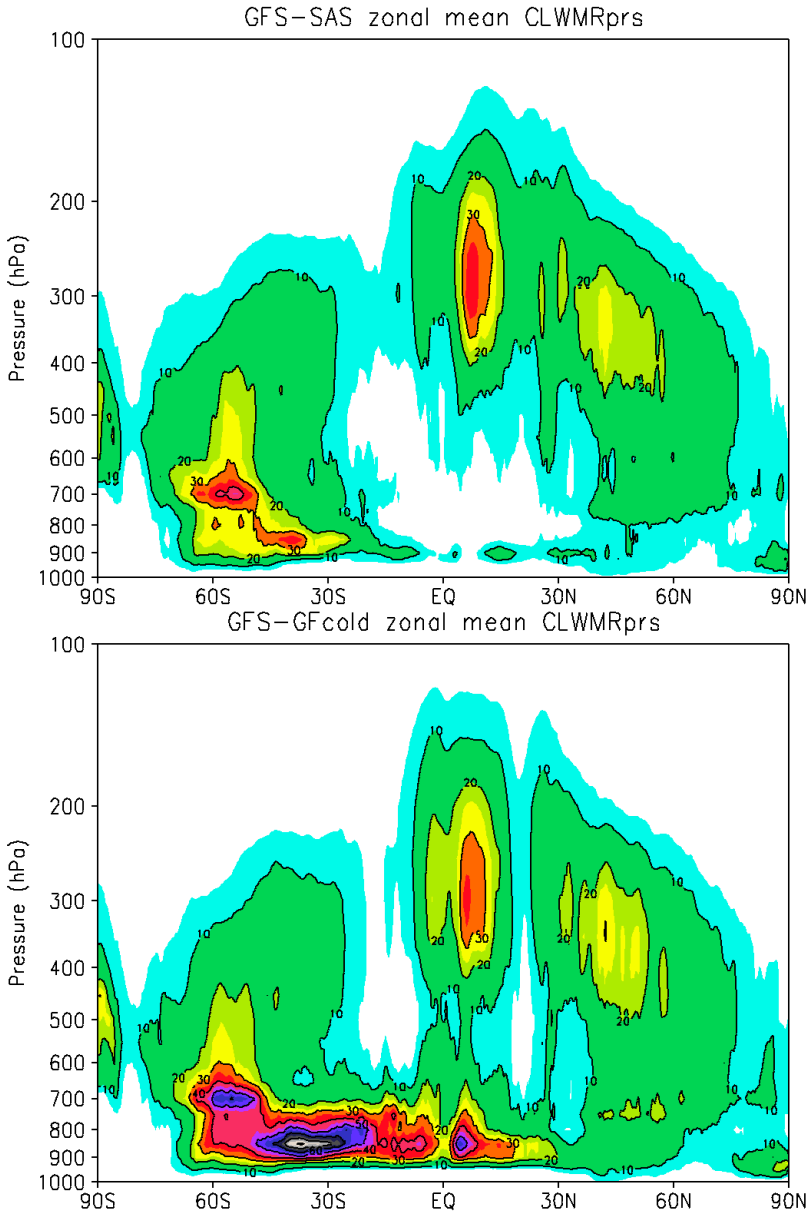


Figure 25. Time series of global mean precipitation rate (mm day^{-1}) for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016, for (a) total, (b) convective, and (c) explicit precipitation rate Occurrence frequency of total precipitation rates between 50°S-50°N is shown in (d). Colors indicate GFS-SAS (red), GFS-GF (blue), GFS-GFcold (green), and observations (black). The black solid and dotted lines denotes the occurrence frequency of the 3-10 day averaged CMORPH and TMPA observations, respectively.

GD_KF3. GFS-GF has more low clouds in the SH and Tropics over the ocean, which leads to a substantially different radiation budget.

As explicit precipitation dominates in GFS-GF, we also expect various cloud properties to change. Condensate mixing ratio (solid and liquid phase clouds) is larger below ~650 hPa in GFS-GF than in GFS-SAS (Fig. 26) in the SH and TROP, which is reflected in the column-integrated condensate mixing ratio (Fig. 27a), as well as in greater PBL, low-level and convective cloud fractions (Fig. 27b, c, f). GFS-GF produces excessive low-level cloud fractions in the ITCZ, SPCZ, Indian Ocean, and South Atlantic Ocean (Fig. 27c). It is also noted that compared to GFS-SAS, GFS-GF suppress the development of parameterized convection over the Amazon basin and produces a smaller high-cloud fraction (Fig. 27e,f).



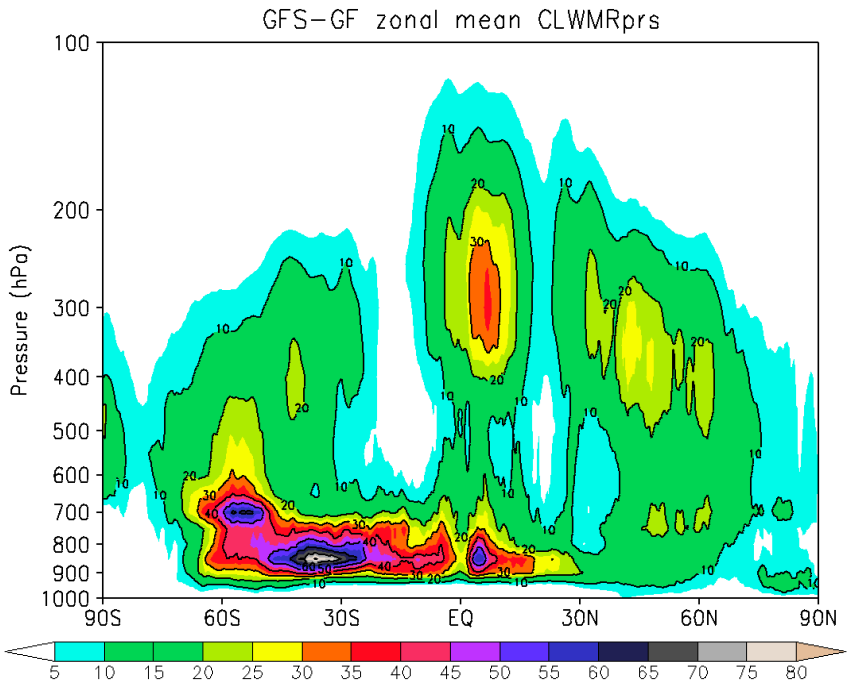
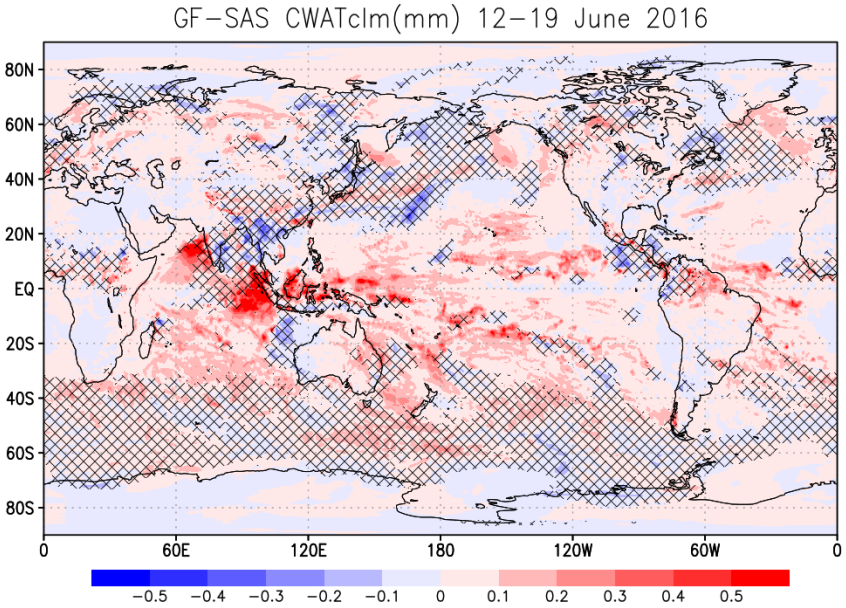
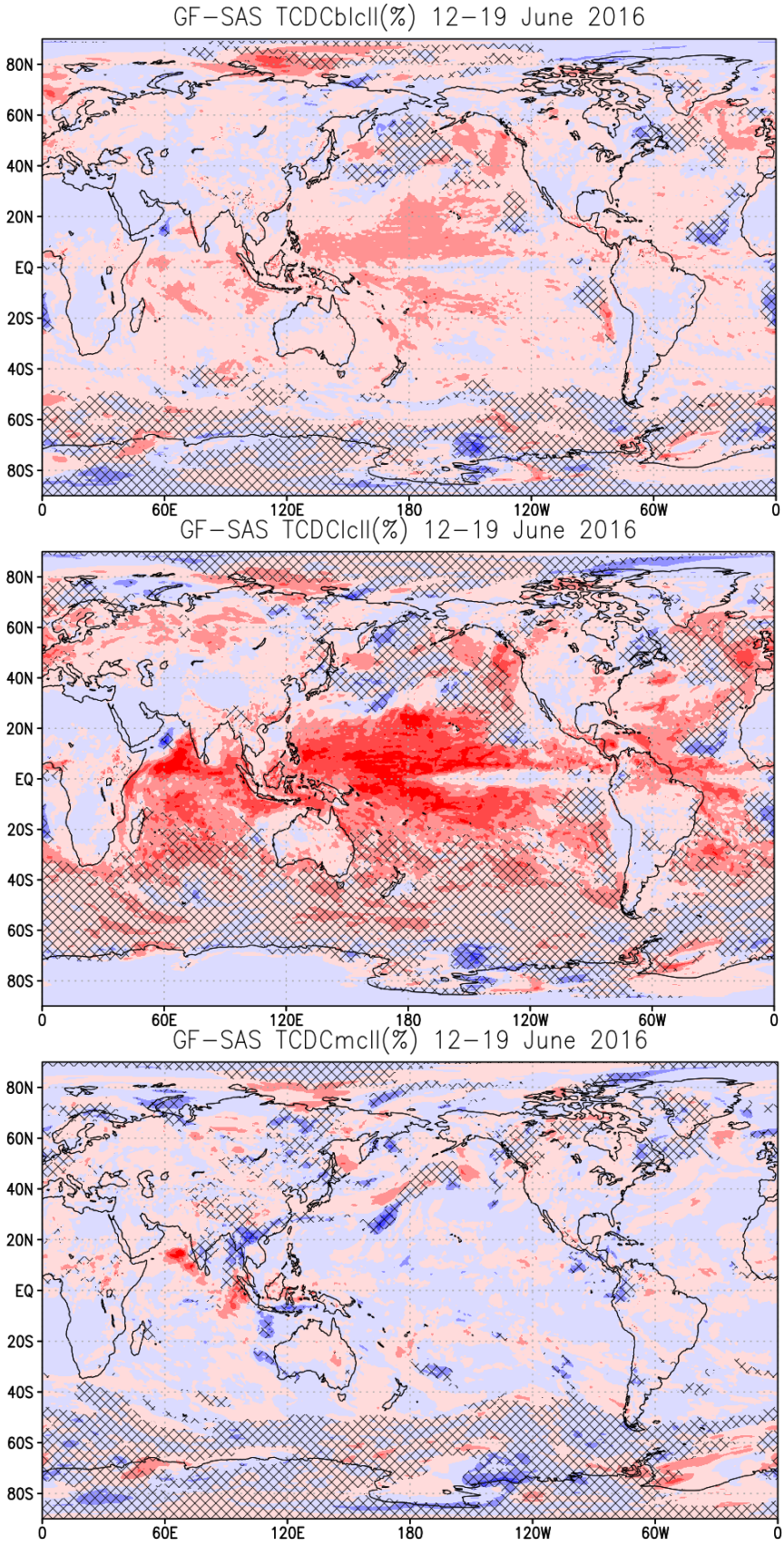


Figure 26. Latitude-pressure plot of the zonal mean of cloud mixing ratio (10^3g kg^{-1}) for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016 in (a) GFS-SAS; (b) GFS-GFcold; (c) GFS-GF.




GF−SAS TCDCblcll(%) 12−19 June 2016
GF−SAS TCDClcll(%) 12−19 June 2016
GF−SAS TCDCmcll(%) 12−19 June 2016
80N
60N
40N
20N
EQ
20S
40S
60S
80S
0
60E
120E
180
120W
60W

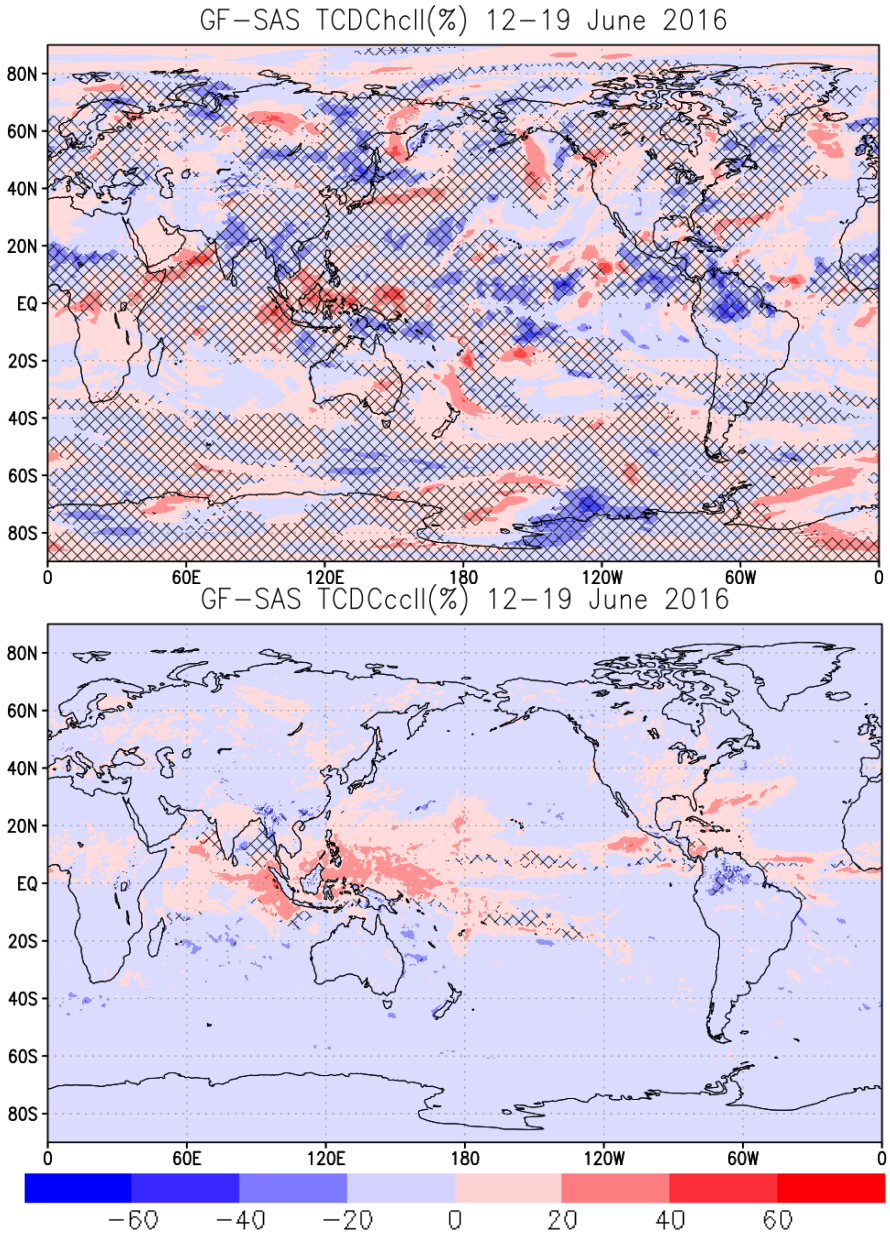


Figure 27. (a) Geographical distribution of the vertically integrate condensate (mm) for 3-10 day forecast initialized at 00 UTC on 10 June, 2016. The forecast cloud water greater than 0.1 mm in GFS-SAS is filled as hatch, the cloud water difference of GFS-GF minus GFS-SAS is shaded in blue and red. Geographical distribution of cloud cover (%) for (b) PBL cloud layer; (c) low-level cloud layer; (d) middle cloud layer; (e) high cloud layer; (f) convective cloud for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016. The model cloud cover greater than 30% in GFS-SAS is filled as hatch; the cloud cover difference of GFS-GF minus GFS-SAS is shaded in blue and red.

The cloud fraction has direct impact on radiation. For example, the change in cumulus parameterization from SAS to GF causes an increase of ~20 Wm^{-2} in upward shortwave (SW) radiation flux at the top of the atmosphere and a decrease of 20 Wm^{-2} in the downward SW radiation flux at the surface (Fig. 28a,e) between 60°S and 60°N. Given that all other terms of the radiative balance are similar between the configurations (Fig. 28d,f), the GFS-GF retains less radiation in the Earth system, which could lead to a different climate than the GFS-SAS.

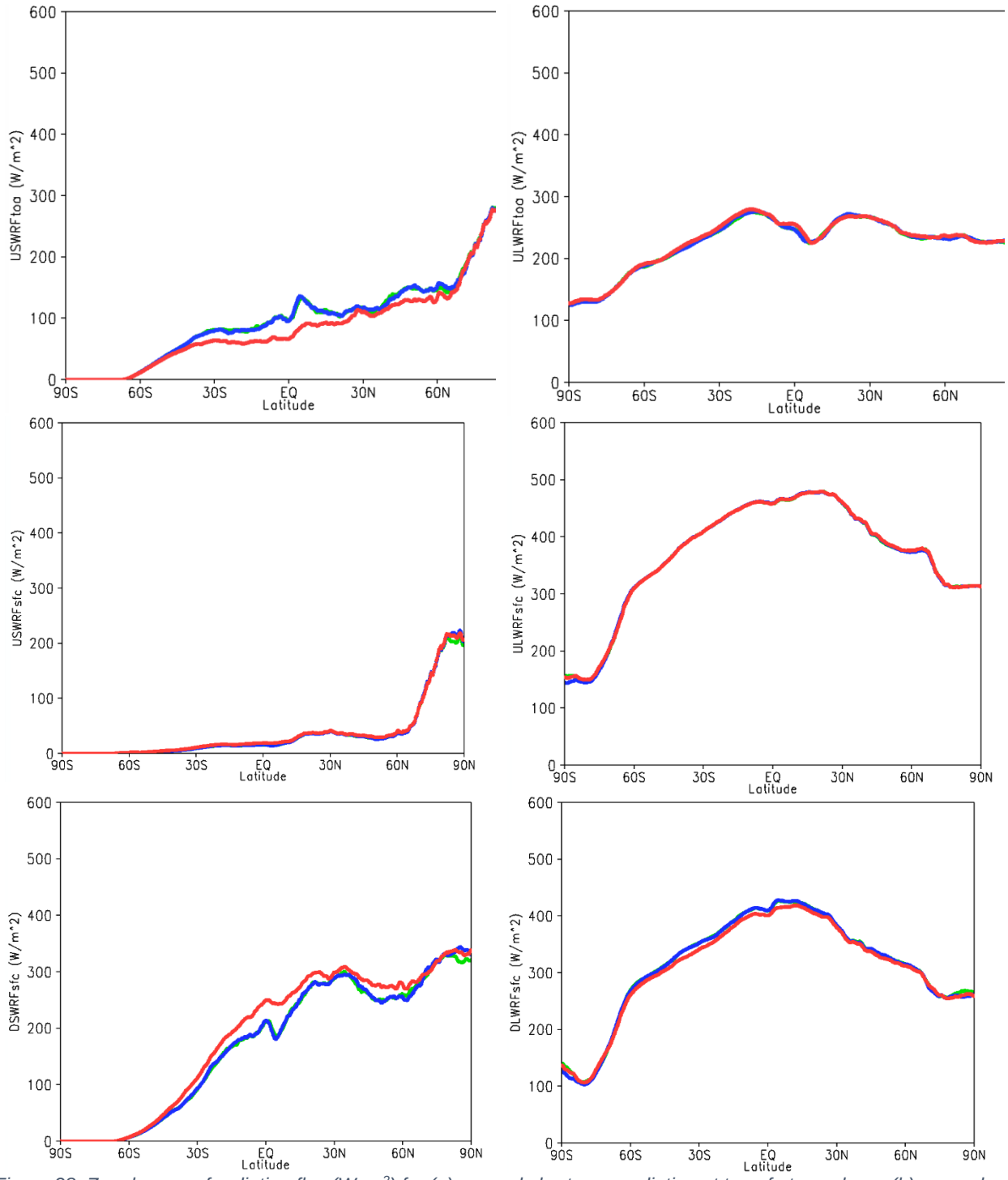


Figure 28. Zonal mean of radiation flux (W m^{-2}) for (a) upward shortwave radiation at top of atmosphere, (b) upward longwave radiation at top of atmosphere, (c) upward shortwave radiation at surface, (d) upward longwave radiation at surface, (e) downward shortwave radiation at surface, and (f) downward longwave radiation at surface for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016. GFS-SAS is displayed in red, GFS-GFcold in green, and GFS-GF in blue.

GD_KF4. The terms of the water budget are different between the GFS-SAS and GFS-GF, with the GFS-GF displaying higher precipitable water.

The time series of terms of global water budget (Fig. 29) shows that precipitation and evaporation are not in balance globally in all configurations. Up to eight days into the forecast, all configurations precipitate less than evaporate, resulting in an overall trend towards increasing PW. The rate of PW increase in GFS-GF is approximately double the one in GFS-SAS.

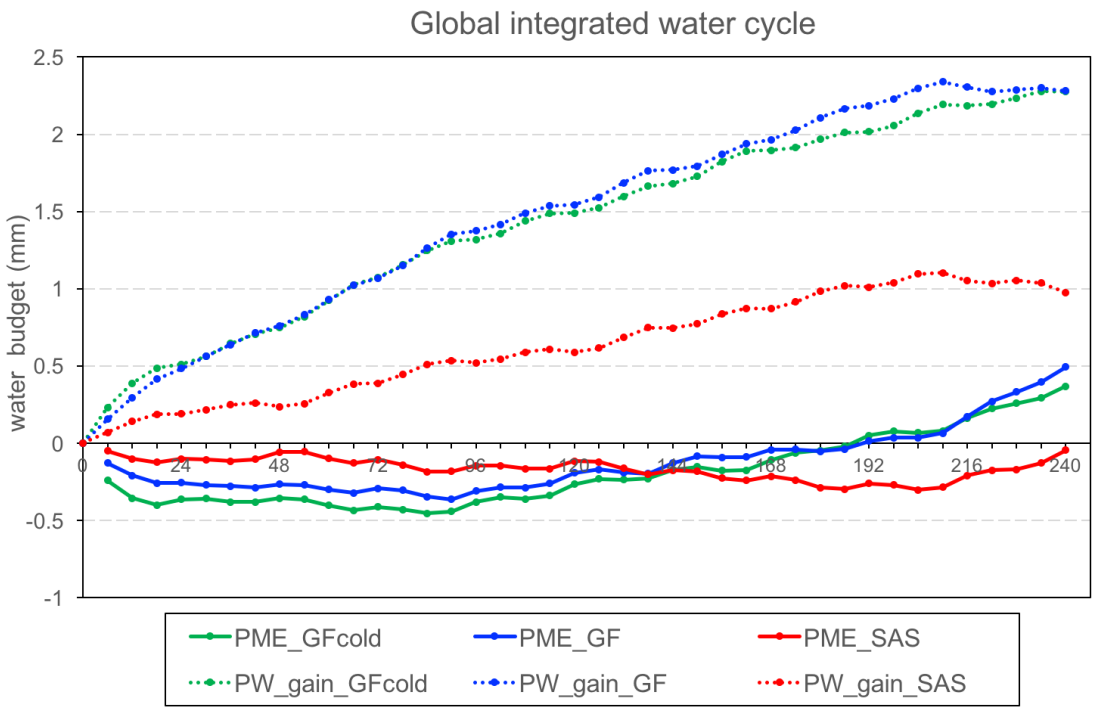


Figure 29. Time series of global water budget terms (mm) precipitation minus evaporation (PME; solid lines) and precipitable water gain (PW_gain; dot-dashed lines) for the 3-10 day forecast initialized at 00 UTC on 10 June, 2016. GFS-SAS is displayed in red, GFS-GFcold in green, and GFS-GF in blue.

Key Findings from Global Verification (GV_KF) of Retrospective Runs

GV_KF1. There is little difference in results between the cold and cycled runs with GF.

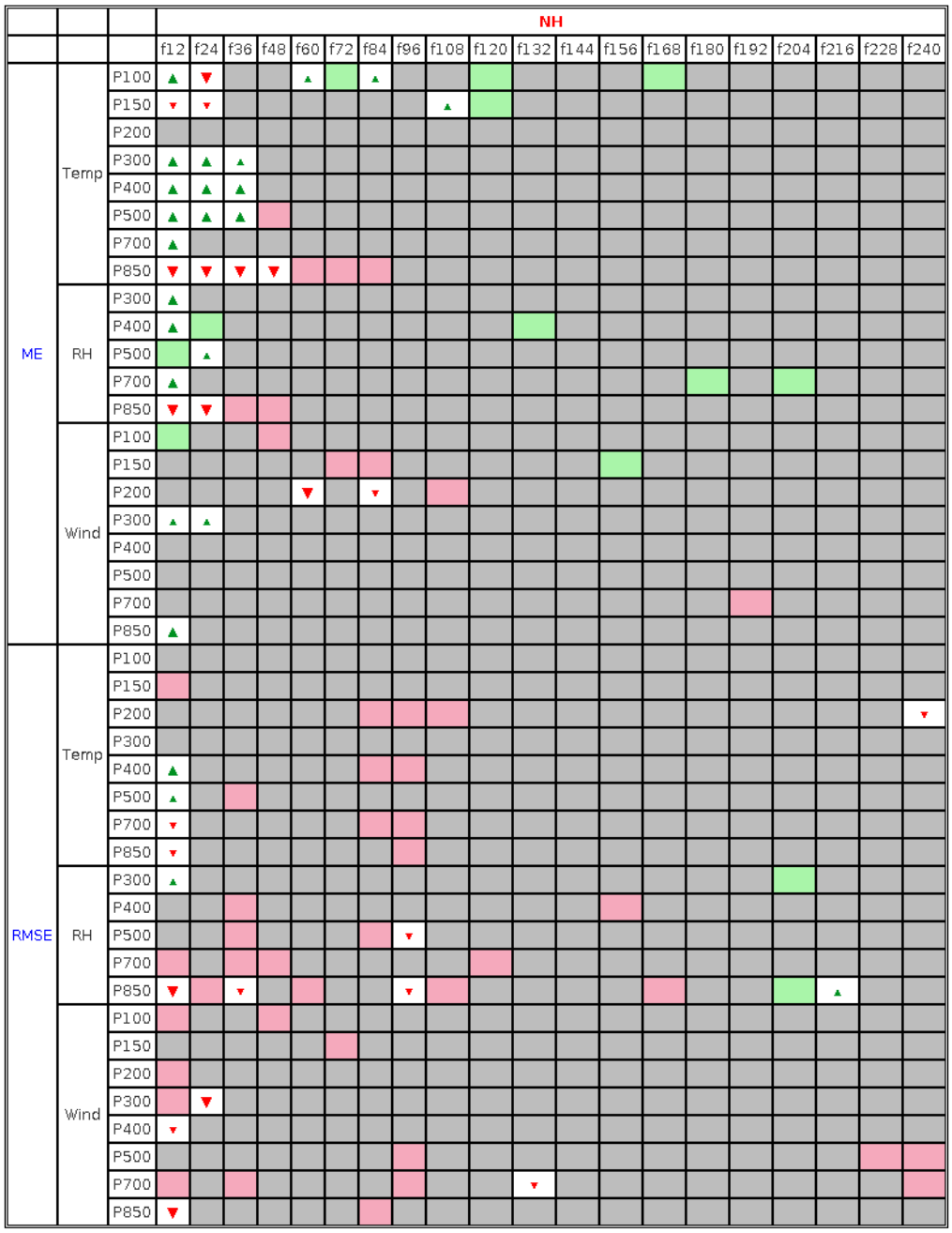
The scorecards show few statistically significance differences between the GFS-GF and GFS-GFcold forecasts. This is exemplified by the upper air scorecard for the NH (Fig. 30a), which is representative of most domains and variables, including precipitation. When present, differences occur early in the forecast, as expected due to the use of different ICs, and are not retained for more than one or two days. This is consistent with the faster spin up of precipitation for GFS-GFcold noted in GD_KF2.

The largest sensitivity to cycling is exhibited in the TROP (Fig. 30b) for temperature bias, but the results are mixed regarding the favored configuration. At upper levels, cycling is beneficial but at 300-400 hPa, it actually degrades the forecast. As further described in GV_KF8, the GFS-GF is more cyclogenetic than GFS-GFcold.

Scorecard

for gfda_0p25_G3 and gfff_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00

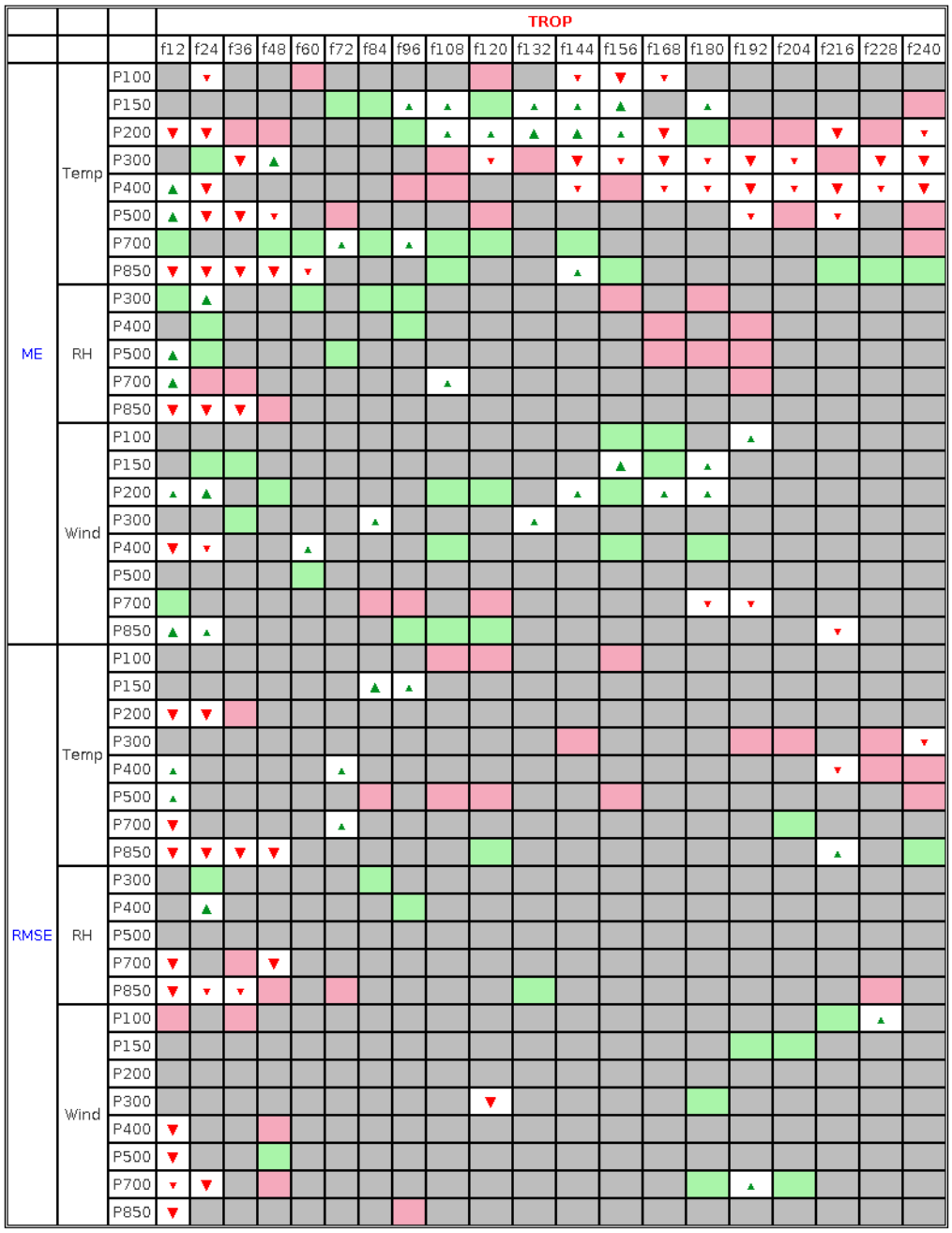


▲	gfda_0p25_G3 is better than gfff_0p25_G3 at the 99.9% significance level
▲	gfda_0p25_G3 is better than gfff_0p25_G3 at the 99% significance level
	gfda_0p25_G3 is better than gfff_0p25_G3 at the 95% significance level
	No statistically significant difference between gfda_0p25_G3 and gfff_0p25_G3
	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 95% significance level
▼	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 99% significance level
▼	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 99.9% significance level
	Not statistically relevant

Scorecard

for gfda_0p25_G3 and gfff_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00



▲	gfda_0p25_G3 is better than gfff_0p25_G3 at the 99.9% significance level
▲	gfda_0p25_G3 is better than gfff_0p25_G3 at the 99% significance level
	gfda_0p25_G3 is better than gfff_0p25_G3 at the 95% significance level
	No statistically significant difference between gfda_0p25_G3 and gfff_0p25_G3
	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 95% significance level
▼	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 99% significance level
▼	gfda_0p25_G3 is worse than gfff_0p25_G3 at the 99.9% significance level
	Not statistically relevant

Figure 30. Scorecard documenting performance of GFS-GF and GFS-GFcold over the a) NH and b) TROP of mean bias and RMSE for temperature (°C , relative humidity (%), and wind speed (m s^{-1}) by forecast lead time and vertical level for June 2-15, 2016. Green (red) marks indicate GFS-GF (GFS-GFcold) is better than GFS-GFcold (GFS-GF). Statistical significance is represented by the type of marks: shading, small arrows, and large arrows indicate 95%, 99%, and 99.9% significance, respectively.

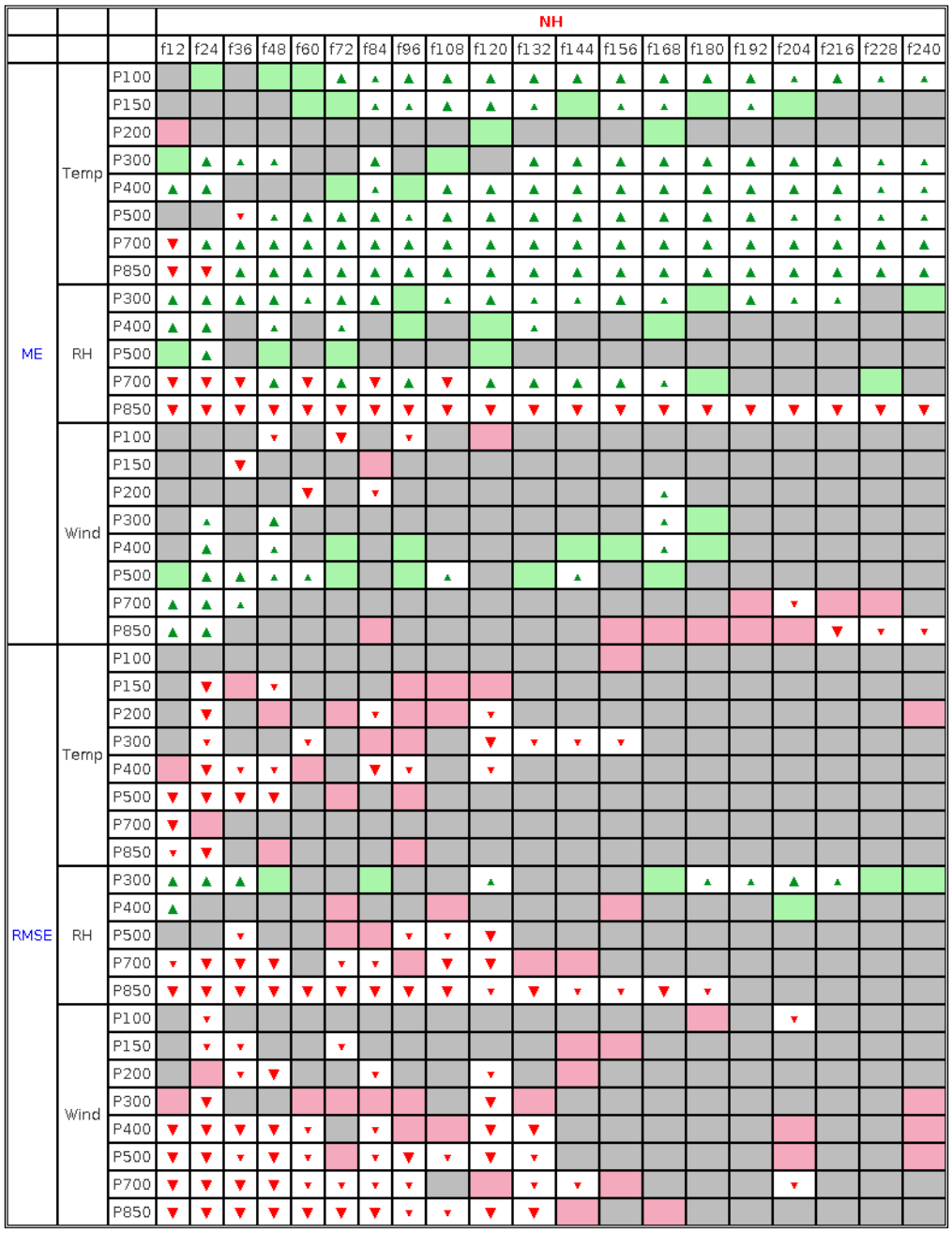
GV_KF2: For most variables and forecast lead times, regardless of global sub-region, GFS-SAS has less RMSE than GFS-GF. The fewest number of differences are noted in the SH, while the most are seen in the TROP region.

The scorecards for the global sub-regions help identify overall patterns in the difference of performance between GFS-SAS and GFS-GF (Fig. 31a-c). The fewest number of SS pairwise differences are noted in the SH (Fig. 31c); a majority of time the pairwise differences are not SS, regardless of forecast variable, level, or lead time. The largest number of SS pairwise differences are noted for the TROP region (Fig. 31b). Within TROP, the GFS-SAS has SS lower RMSE compared to GFS-GF at a majority of forecast levels and lead times for temperature. In addition, for wind speed, most forecast levels and lead times indicate a SS pairwise difference that favors GFS-SAS during the first five days. There are far fewer SS pairwise differences beyond day five for wind speed; however, it is worth noting that late in the forecast period several SS pairwise differences are noted at 100 hPa that favor the GFS-GF. For RH, a majority of the SS pairwise differences favor the GFS-SAS at all except the highest levels examined, where GFS-GF is preferred. For the NH, GFS-SAS continues to be most frequently favored in terms of RMSE; however, there is an overall reduction in the number of SS pairwise differences for all forecast variables, levels and lead times (Fig. 31a).

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00

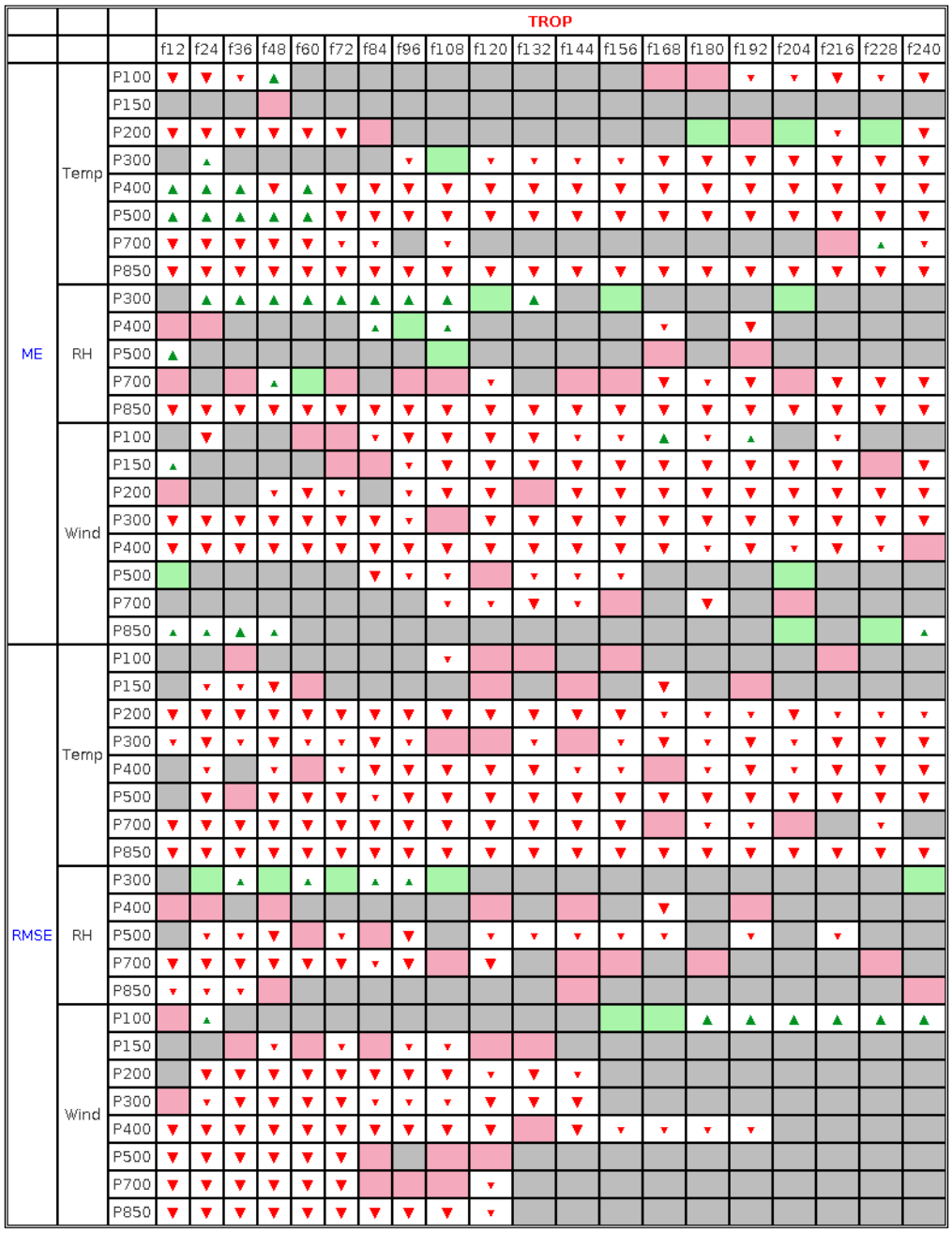


▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
	gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
	No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
	Not statistically relevant

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00

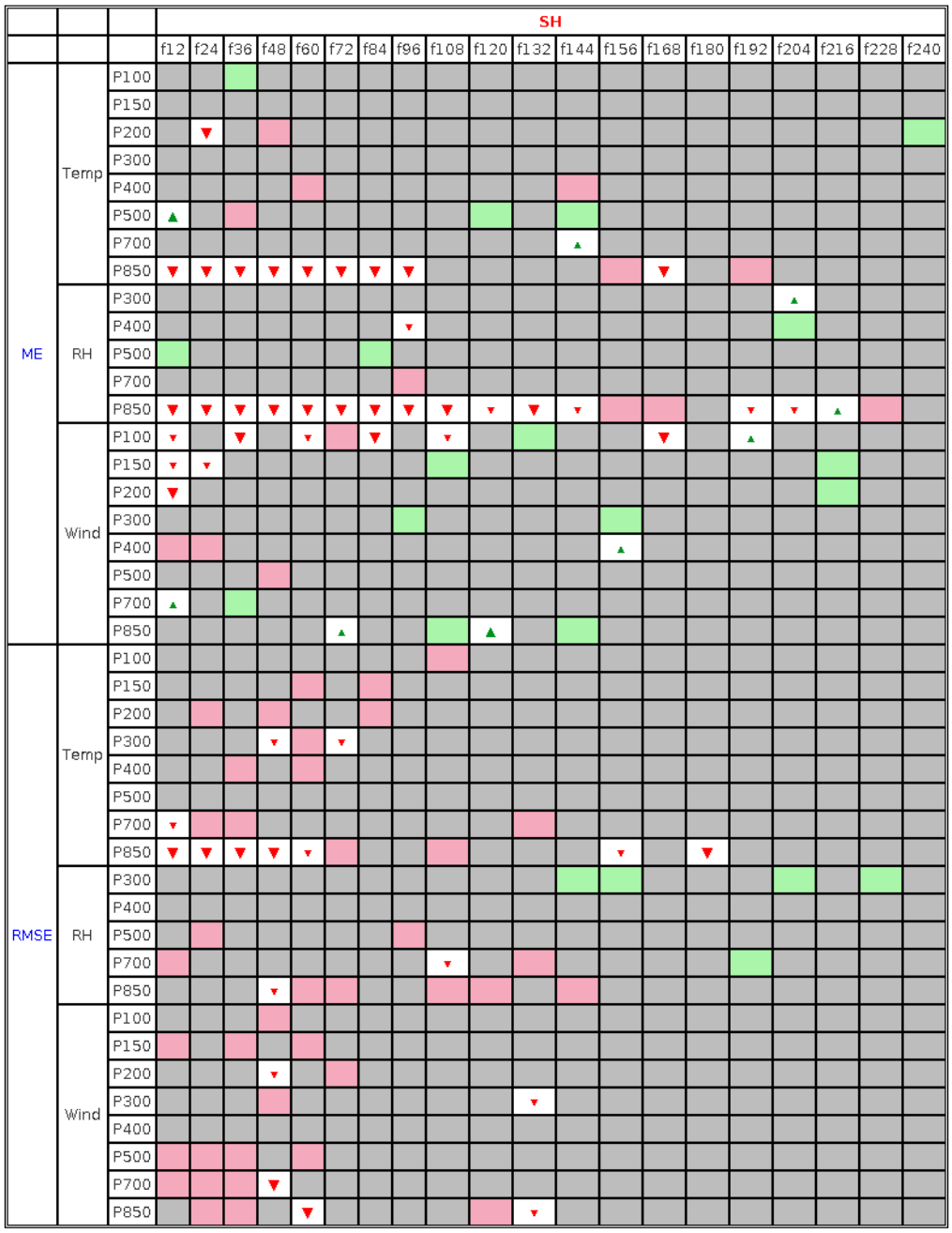


▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
	gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
	No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
	Not statistically relevant

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00



▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
▲	gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
	gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
	No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
▼	gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
	Not statistically relevant

Figure 31. Scorecard documenting performance of GFS-GF and GFS-SAS over the a) NH, b) SH, and c) TROP of mean bias and RMSE for temperature (°C), relative humidity (%), and wind speed (m s^{-1}) by forecast lead time and vertical level for June 2-15, 2016. Green (red) marks indicate GFS-GF (GFS-SAS) is better than GFS-SAS (GFS-GF). Statistical significance is represented by the type of marks: shading, small arrows, and large arrows indicate 95%, 99%, and 99.9% significance, respectively.

GV_KF3: The profiles of temperature bias are different between the GFS-SAS and GFS-GF with the better performer depending strongly on sub-region; the GFS-GF is preferred over the NH and GFS-SAS is generally preferred for the TROP.

When examining the overall scorecard results for upper-air temperature bias aggregated by level and forecast lead time over the NH sub-region (Fig. 32a), the GFS-GF has SS less bias in the low-to-mid and upper-most levels of the atmosphere with a few exceptions early in the forecast period. When looking at two particular forecast lead times (Fig. 32a; day 5 and 10 shown), both configurations exhibit a warm bias in the low-to-mid levels of the atmosphere transitioning to a cold bias at the upper levels. The bias values generally increase as forecast lead time increases for GFS-SAS and GFS-GF.

In contrast, over the TROP sub-region, GFS-SAS has SS lower bias values in the low-to-mid levels (most consistently at 850 and 500-300 hPa), with a few exceptions in the mid-levels throughout the forecast period (Fig. 32b). Looking at the day 5 and 10 temperature bias profiles (Fig. 32b), the GFS-SAS has a relatively small cold bias throughout the column, whereas the GFS-GF exhibits a larger cold bias near the surface, transitioning to a warm bias in the mid-troposphere and back to a cold bias at the upper-most level. The cold bias near the surface in GFS-GF may be related to the excessive low-level cloudiness discussed in GD_KF3. The biases generally increase more with forecast lead time for GFS-GF compared to GFS-SAS.

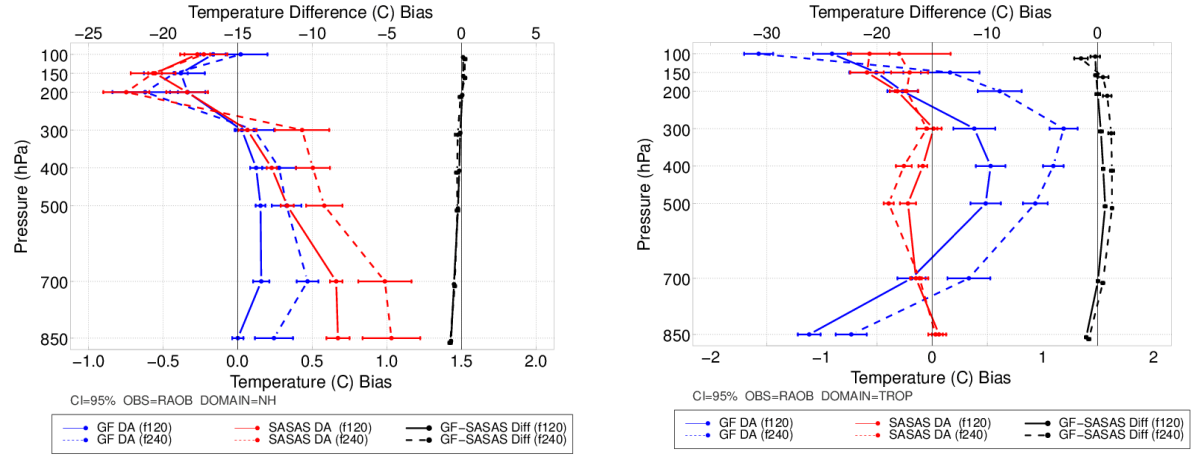


Figure 32. Vertical profile of the median bias for temperature (°C) aggregated for June 2-15, 2016 over the a) NH and b) TROP regions. The 120-h forecast lead time is represented by the solid lines and the 240-h forecast lead time is dashed. GFS-GF is blue, GFS-SAS is red, and the difference (GFS-GF minus GFS-SAS) is black. The horizontal bars surrounding the aggregate value represent the 95% CIs.

GV_KF4. GFS-SAS is warmer than GFS-GF over the CONUS at 2m, and the two configurations have distinct diurnal cycle of errors: GFS-SAS warms up too quickly in the daytime, while GFS-GF maximum temperatures are below observed. A problem noted in a previous GMTB test using the FY16 GFS

Physics suite, of CONUS 2-m temperatures increasing with forecast lead time in GFS-SAS runs, has not been seen in this test.

Both GFS-GF and GFS-SAS have distinct diurnal cycle of errors when considering 2-m temperature over the CONUS (Fig. 33). Both configurations have a minimum cold-to-neutral bias at 00 UTC, with GFS-GF typically colder than both GFS-SAS and the observations, perhaps indicating GFS-GF has difficulty maintaining peak daytime temperatures. Also notable is a difference in phase shift with regards to maximum temperatures. The GF-SAS warms up too quickly during the daytime, with maximum biases occurring around 18 UTC, while GFS-GF has smaller peak bias values that are shifted 6 hours prior. In general, due to GFS-GF having smaller maximum daytime temperatures than GFS-SAS, GFS-GF is often closer to an unbiased forecast at these times. In conjunction with the diurnal cycle of errors between the two configurations, SS pairwise differences also follow a diurnal cycle, with GFS-GF better at 18 UTC and GFS-SAS better at 00 UTC. Due to the 14-day sample, it is important to keep these results in context as the CIs are often wide and encompass 0, indicating an unbiased forecast.

It is also noted that in a previous non-cycled test conducted by GMTB employing the FY16 GFS-SAS, the 2-m temperature over the CONUS displayed a gentle increase as the forecast lead time progressed (Fig. 34). This undesirable behavior was not seen in this test using the FY17 GFS physics using the scale-aware SAS.

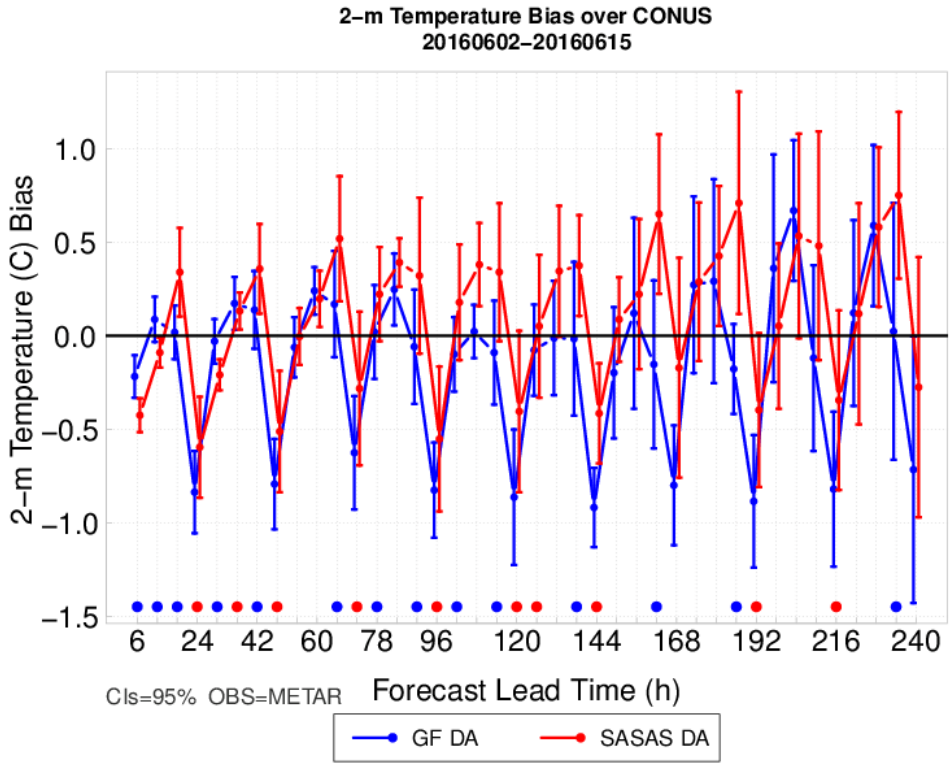


Figure 33. Two-meter temperature bias (°C) over the CONUS domain versus forecast lead time (h) for GFS-SAS (red) and GFS-GF (blue) for June 2-15, 2016. The dots across the bottom indicated SS pairwise differences; the blue (red) dots indicate GFS-GF (GFS-SAS) is better at the 95% significance level.

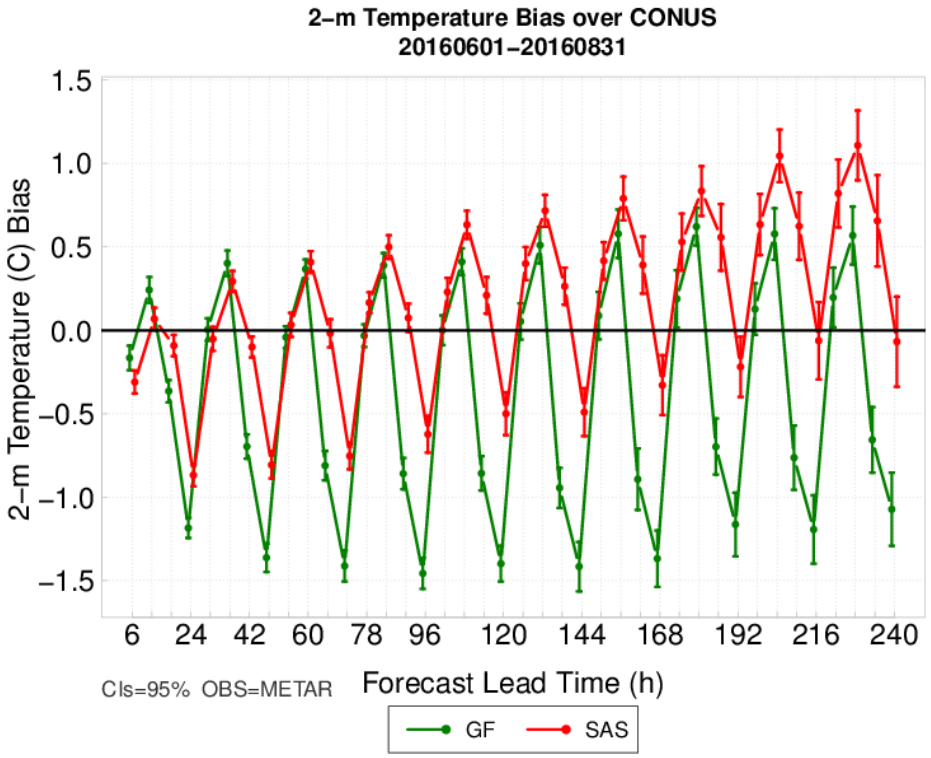


Figure 34. Two-meter temperature bias (°C) over the CONUS domain versus forecast lead time (h) for GFS-SAS (red) and GFS-GFcold (green) from the previous uncycled GF Test for June 1 - August 31, 2016.

GV_KF5: Wind biases are similar between the GFS-SAS and GFS-GF throughout the atmosphere in the NH and SH, but the GFS-GF has larger negative biases in the TROP sub-region, especially at upper levels.

The scorecards for wind speed bias as a function of level and forecast lead time highlight relatively few SS pairwise differences between the two configurations for the NH (Fig. 31a) and SH (Fig. 31c) sub-regions with significantly more differences favoring GFS-SAS noted for the TROP sub-region (Fig. 31b). These SS pairwise differences are generally noted for the mid-to-upper levels of the atmosphere. While there are fewer SS pairwise differences between the two configurations over the NH sub-region, both the GFS-SAS and GFS-GF exhibit a low wind speed bias at all levels with similar bias values at both day 5 and 10 (Fig. 35a). Similarly, Fig. 35b shows wind speed bias as a function of pressure level over the TROP sub-region for the 5- and 10-day forecasts. While the GFS-SAS generally remains relatively unbiased for most levels, the GFS-GF has a clear low wind speed bias, in particular for the mid-to-upper levels of the atmosphere.

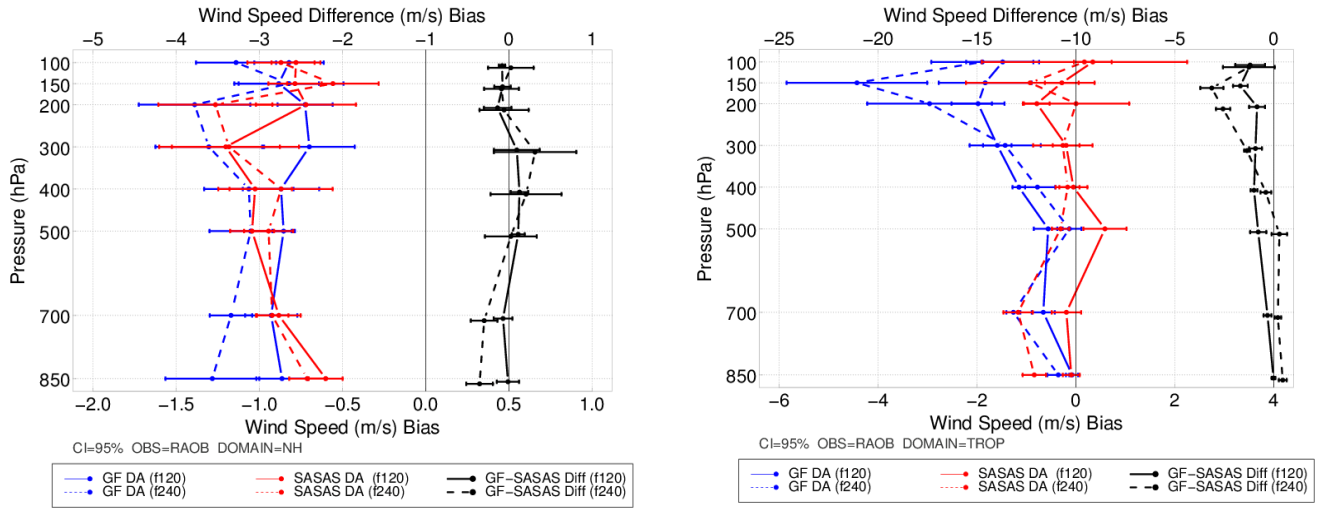


Figure 35. Same as Fig. 33, except for wind speed (m s^{-1}).

GV_KF6. Precipitation placement is better in GFS-SAS than GFS-GF.

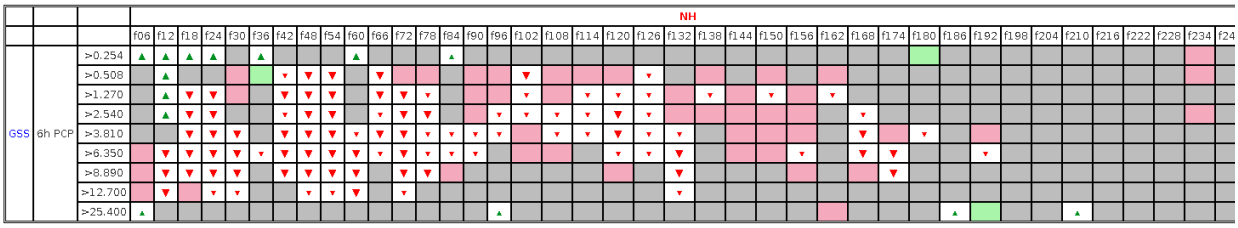
When considering the 6-h accumulation ETS across global regions, several overarching results emerge. Overall, GFS-SAS has more skill than GFS-GF (Fig. 36a-d). This is amplified between the 0.508 mm and 12.700 mm thresholds and at earlier to middle forecast lead times. The most differences favoring GFS-SAS are in the tropical region, while the least amount of differences favoring GFS-SAS occur in the SH and Amazon regions. While GFS-SAS has overall better precipitation placement than GFS-GF, there are several notable exceptions. When GFS-GF outperforms GFS-SAS, it is typically at the lowest (0.254 mm) and highest (25.4 mm) thresholds. In addition, in the SH, GFS-GF has SS higher ETS at a number of lead times at the 0.254 mm threshold and several of early lead times at the 0.508 mm threshold. Over CONUS, when there are SS differences, GFS-SAS almost always has higher skill than GFS-GF (Fig. 36e).

The daily precipitation ETS scorecards have similar results (not shown). While there are a number of no SS differences across forecast lead time and threshold, when there are SS differences, GFS-SAS typically has better placement than GFS-GF. These SS differences often occur at earlier forecast lead times and lower thresholds.

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00

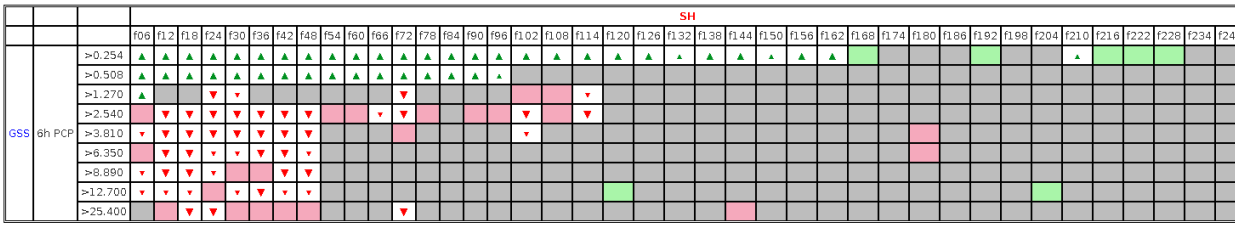


gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
Not statistically relevant

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00

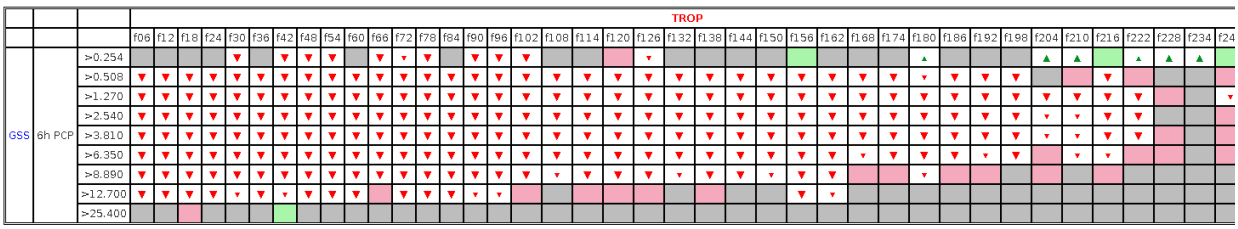


gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
Not statistically relevant

Scorecard

for gfda_0p25_G3 and sasda_0p25_G3

2016-06-02 00:00:00 - 2016-06-15 00:00:00



gfda_0p25_G3 is better than sasda_0p25_G3 at the 99.9% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is better than sasda_0p25_G3 at the 95% significance level
No statistically significant difference between gfda_0p25_G3 and sasda_0p25_G3
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 95% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99% significance level
gfda_0p25_G3 is worse than sasda_0p25_G3 at the 99.9% significance level
Not statistically relevant

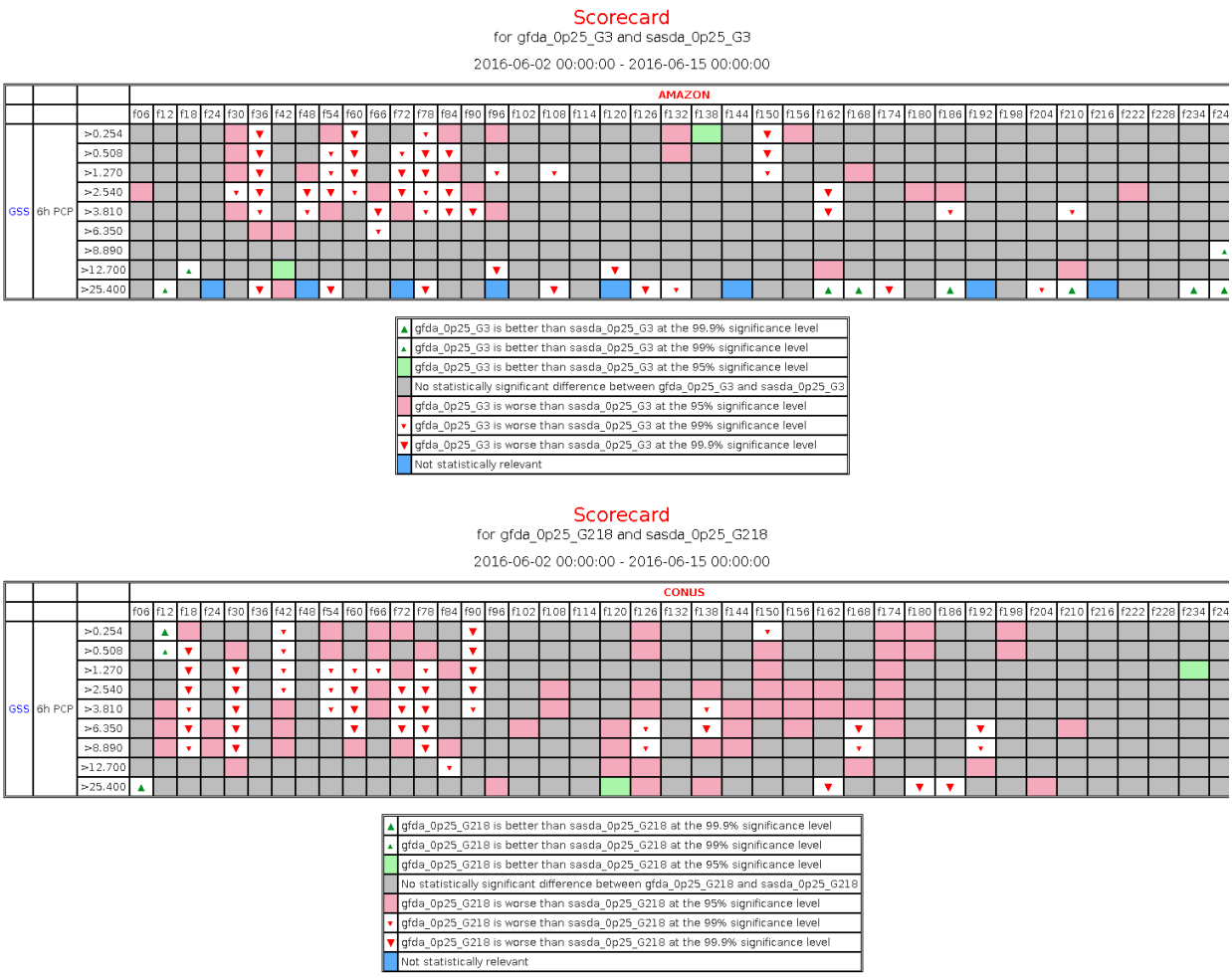


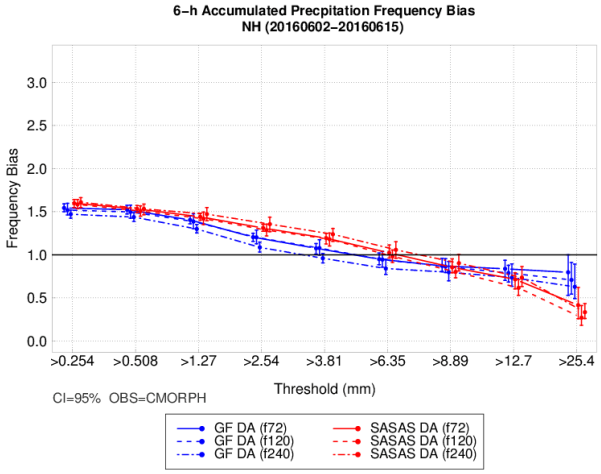
Figure 36. Scorecard documenting performance of GFS-GF and GFS-SAS over the a) NH, b) SH, c) TROP, d) Amazon, and e) CONUS of aggregate ETS for 6-h accumulated precipitation by forecast lead time and precipitation threshold for June 2-15, 2016. Green (red) marks indicate GFS-GF (GFS-SAS) is better than GFS-SAS (GFS-GF). Statistical significance is represented by the type of marks: shading, small arrows, and large arrows indicate 95%, 99%, and 99.9% significance, respectively.

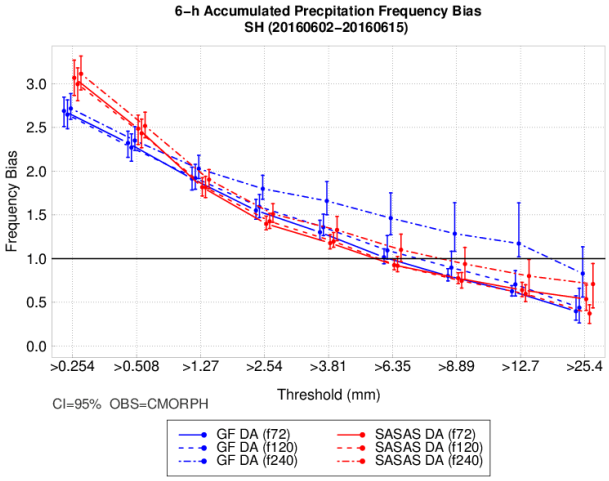
GV_KF7. The configuration that predicts better precipitation coverage depends on domain. Both configurations see an increase in precipitation coverage with forecast lead time.

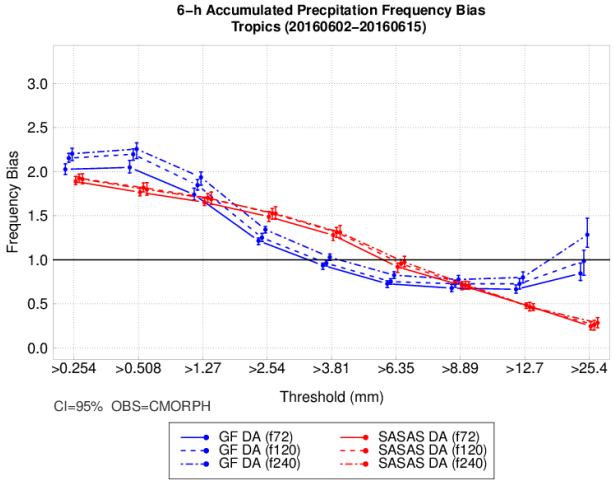
The GFS-SAS and GFS-GF display differing performance with regard to 6-hr accumulated precipitation coverage amount depending on domain (Fig. 37a-e). However, one commonality seen for both configurations in all regions with exception of the NH, is that with increasing forecast lead time, an increase in frequency bias is seen for most thresholds. This was also illustrated in the precipitation time series shown in the diagnostic analysis (Fig. 25a).

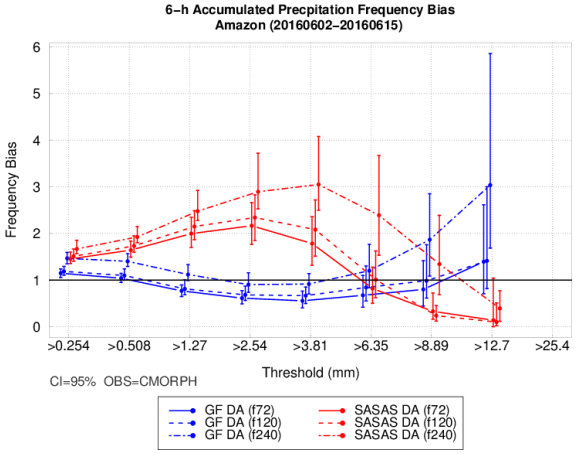
In the NH, SH, and tropical regions for the lead times shown, both configurations have a modest over-forecast up to the 2.54 mm threshold, followed by a general decrease in frequency bias as threshold increases. In the NH, GFS-GF better predicts precipitation coverage at all thresholds (Fig. 37a), while in the SH, GFS-GF is superior only the lowest thresholds (Fig. 37b). Mixed performance is seen in the TROP, where GFS-GF has excessive coverage when compared to GFS-SAS at the lowest thresholds, but generally

has more unbiased forecasts at several middle and higher thresholds (Fig. 37c). In the land-dominated Amazon region, the frequency bias curves for the two configurations differ more than on the other global regions (Fig. 37d) with GFS-GF having a near-neutral bias at low-to-middle thresholds, while GFS-SAS has a SS high bias at all forecast hours up to the 3.81 mm threshold. In the CONUS, GFS-SAS at the 3- and 5-day forecasts is generally more unbiased than GFS-GF. In the Amazon and CONUS regions, the trend of increasing bias with forecast lead time is accentuated, most notably for GFS-GF at the highest thresholds in the Amazon as well as GFS-SAS at most thresholds in both land-dominated regions.









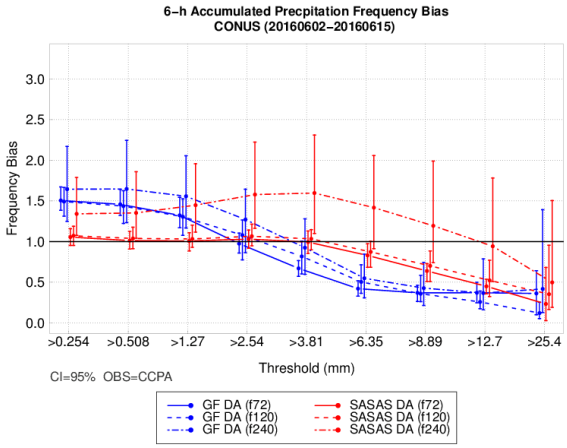
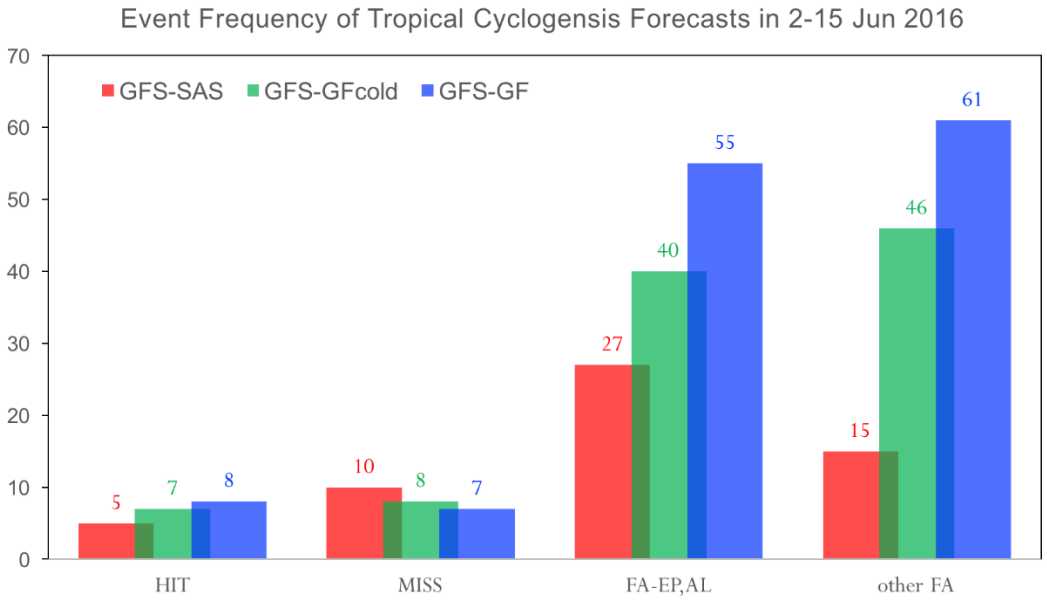
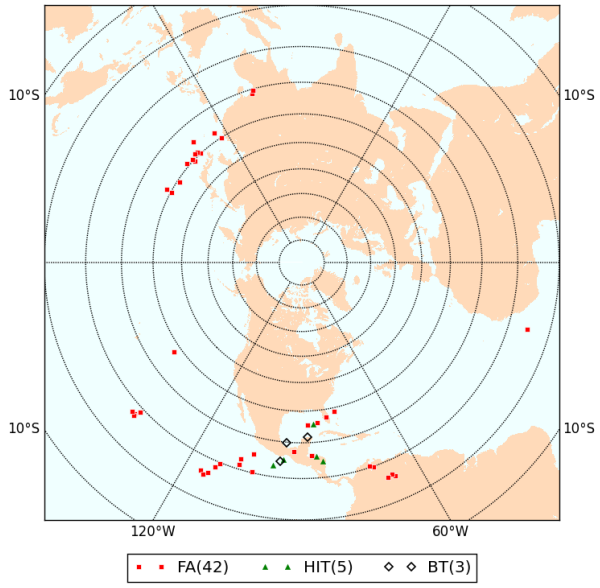


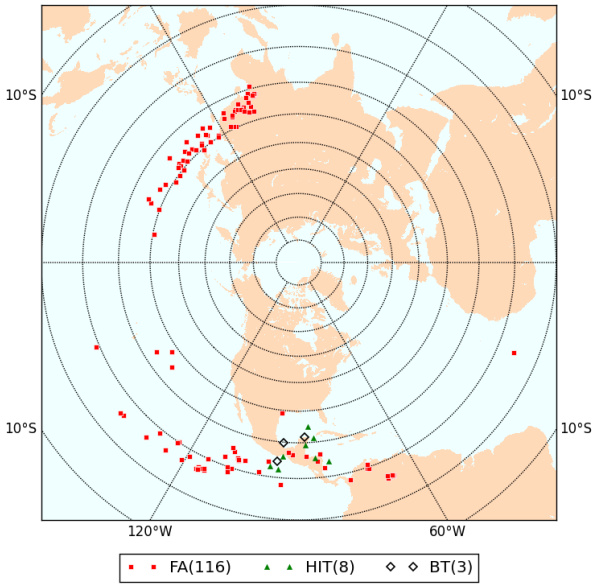
Figure 37. Frequency bias of 6-h accumulated precipitation (mm) for GFS-SAS (red) and GFS-GF (blue) aggregated over the a) NH, b) SH, c) tropical region, d) Amazon, and e) CONUS for June 2-15, 2016. The 72-h forecast lead time is represented by the solid lines, the 120-h forecast lead time in dashed, and the 240-h forecast lead time is dot-dashed. The vertical bars surrounding the aggregate value represent the 95% CIs.

GV_KF8. GFS-GF is more cyclogenetic and produces more tropical cyclogenesis false alarms than GFS-SAS.

Model data are available during 2-25 June 2016, and during this period, only three tropical cyclogenesis events were observed in the BT data: TS Colin (AL03, 21.6°N, 88.0°W) at 12 UTC June 5, TS Danielle (AL04, 19.9°N, 94.7°W) at 18 UTC June 19 and TD One-E (EP01, 14.4°N, 96.3°W) at 00 UTC June 7. The 0000 UTC initializations of each model were analyzed with a forecast window of 240 h. Thus, a perfect model would have 15 hits, predicting AL03, AL04 and EP01 in 4, 6, and 5 initializations, respectively. GFS-SAS, GFS-GF, and GFS-cold produced 47, 124, and 93 TCs, respectively, but the number of hits did not reach 15 (Fig. 38a). The large number of cyclogenesis led to numerous false alarms in all configurations, with more serious problems in the configurations using GF, especially for the WP basin (Fig. 38a-d). It is interesting that the false alarms were exacerbated through the cycled DA, suggesting that cyclogenesis is sensitive to the differences in temperature forecast discussed in KF1.


Event Frequency of Tropical Cyclogensis Forecasts in 2-15 Jun 2016
GFS-SAS
GFS-GFcold
GFS-GF
70
60
50
40
30
20
10
0
5
7
8
10
8
7
27
40
55
15
46
61
HIT
MISS
FA-EP,AL
other FA


10°S
10°S
10°S
10°S
120°W
60°W
FA(42)
HIT(5)
BT(3)


10°S
10°S
10°S
10°S
120°W
60°W
FA(116)
HIT(8)
BT(3)

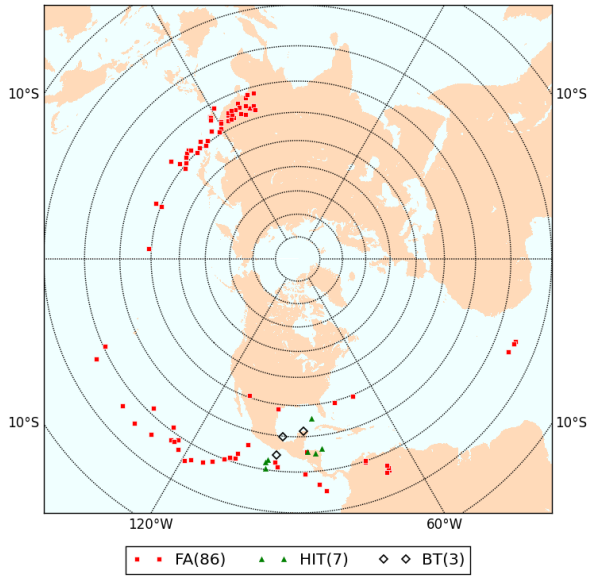


Figure 38. a) Northern Hemisphere tropical cyclogenesis verification scores (hits, misses, false alarms) for GFS-SAS (red), GFS-GFcold (green), and GFS-GF (blue). False alarms are shown separately for the AL/EP basins and all other northern hemisphere basins. Spatial distribution of observed (BT; diamonds) cyclogeneses, hits (green triangles), and false alarms (red dots) for b) GFS-SAS, c) GFS-GF, and d) GFS-GFcold. Genesis locations in the CARQ are not depicted in this plot.

Discussion and Conclusions

This test illustrated the growth in the GMTB testing hierarchy by employing four tiers for assessing the GF: SCM and GFS run in case study mode for diagnostics, multi-day cold starts, and multi-day with cycled DA. It also showed GMTB's agility to quickly upgrade its systems to run a new version of the GFS (FY17) and absorb revised procedures to run the GFS (using the Rocoto Workflow Management System). Moreover, new verification procedures were added, such as scorecards for precipitation and scores for tropical cyclogenesis. With this evolution, the GMTB positions itself as an important contributor for testing and evaluation of physics innovations for NCEP forecast applications.

While the assessment conducted for this report was comprehensive, the GMTB intends to continue enhancing its harness for future tests. The algorithm for identification of tropical cyclogenesis should be reviewed to ascertain whether the physics needs to be tuned or the tracker needs to be adjusted to detected less cyclogeneses, especially over land. One diagnostic under consideration is an assessment of the water budget in GFS experiments with the cumulus parameterization (or even the entire physics suite) disabled, as it appears that neither GFS-SAS or GFS-GF are conservative. Another tool being explored is the use of DA products for the purposes of forecast verification. The GDAS outputs differences between analyses and background (A-B), which are the increment added by the DA system to the prior. These increments can be used to assess biases introduced by the model physics, and can produce more robust results than assessment of the forecast biases because GDAS results are available every six hours, while experimental forecasts are usually only run once a day to conserve computational resources while spanning a longer window of time. As an example, If one only has computational resources to run 14 forecasts, it is best to run two weeks of forecasts initialized once a day than 3.5 days

of forecast initialized every six hours. However, to run 14 forecasts initialized daily, 56 runs of GDAS are needed, yielding a robust sample.

Important collaborations were established in the planning and conducting of this assessment. In order to decide the content and configuration of the test, GMTB reviewed various efforts in physics developments and presented options to EMC and NGGPS representatives, who actively participated in making the decision to test GF in cycled mode. The developer, Georg Grell, was involved early in the planning stages and took advantage of the testbed to conduct both SCM and global runs to perform his own sensitivity experiments to gain understanding of the results and devise further avenues of development. These procedures are consistent with the governance proposed for the Common Community Physics Package (CCPP), a software framework that is under development and was not utilized for this test.

The key findings for this test, summarized in the executive summary and expanded in the body of the report, will not be repeated here for brevity. It suffices to say that there were pronounced differences between forecasts run with SAS and GF, with an overall advantage in the verification scores for GFS-SAS. The hypothesis that the performance of GF forecasts would be better if the ICs were created using a consistent physics suite was refuted, as little difference was seen between cold and cycled runs with GF. The inferiority of GFS-GF when compared to GFS-SAS is not unexpected since the entire physics suite may need to be re-tuned to use GF. This poses a complex question regarding the advancement of the GF scheme to higher tiers of testing and a potential proposal for inclusion in the supported CCPP, making a candidate for R2O. While GF is a mature development, already operational at NCEP in the RAP, and extensively tested in the experimental global Flow-following finite-volume Icosahedral Model (FIM) developed by ESRL, it may take time for it to demonstrate superiority in the GFS. Even then, it is not clear it could perform better than the scale-aware SAS, as it should be noted that there are a lot of similarities between the two schemes, given that both use the mass-flux and that the SAS implementation in GFS is partially derived from Grell (1993).

Given the deadlock created by inferior performance of the untuned GFS-GF shown in this test, GMTB recommends that the GF not be advanced to the next tier of testing at this time. Instead, GMTB recommends that the developer continue to take advantage of the GMTB testing facility to improve the scheme and possibly perform initial tuning. If the developer can demonstrate some encouraging results, future testing by GMTB can be considered. Another avenue for going forward is a new effort funded by NGGPS to analyze differences between the GF and scale-aware SAS schemes, with the goal of producing a joint recommendation of the most appropriate version for the GFS advanced physics suite, which may be a slight or major modification of either scheme. The principal investigator, Georg Grell, and his EMC collaborator, Jongil Han, may also compare the GF and SAS schemes to other cumulus parameterizations, such as the one described in Chikira and Sugiyama (2010). Once their recommendation, and possible revised or combined code is available and preliminarily tested, GMTB could conduct further assessment of the scheme(s).

The GMTB has been conducting its evaluation in an environment of extremely limited computational resources. Its allocation of 100,000 core-hours a month on Theia was completely used up

by the two sets of 14-day tests with once-daily initializations (GFS-SAS and GFS-GF), leaving the GFS-GFcold to run slowly with unallocated resources when the machine was available (windfall queue). For this reason, no experiments were performed to assess the scale-awareness of the schemes, as this would have required expensive runs in higher-resolution. Going forward, it is important that the testbed be provided enough resources to provide relevant results for informing NCEP.

Developer comments

Parameterizations are components of atmospheric computer models that aim to represent the statistical effects of a physical process that cannot be directly treated by a computer model, usually when the spatial resolution is not sufficient to resolve the process directly (e.g., convective clouds). With a myriad of processes handled inside atmospheric computer models, physical parameterizations virtually interact with many other components of a modeling system, and they are a key ingredient for model realism and skill. The GF convective parameterization was built based on a series of convective parameterizations developed by Grell (1993) and expanded by Grell and Devenyi (2002). The 2014 version of GF was the first time that Arakawa's scale awareness ideas were applied successfully for weather prediction. The version that was used in these initial tests with the GFS is based on the 2014 version but has momentum transport included. Since the test conducted by GMTB was considered a baseline test before tuning, some options of the scheme were not turned on. In this section, we will briefly show the sensitivity to tuning, both inside the convective parameterization, but also within the suite of parameterizations. We do not intend to indicate which of the tuned parameters are better or worse, but only get a feeling of the sensitivity to some of the possible tuning constants. This includes the tri-modal application of the scheme (mid-level convection was turned off) and the stochastic application (temporal and spatial correlations can be used to perturb the normalized vertical mass flux PDFs). Another new implementation described in GF, but not used, is the interaction with aerosols.

To demonstrate the impacts of tuning, two additional experiments were conducted using the GMTB workflow for a ten day period, spanning 20160601-20160610; all runs were initialized at 00 UTC, with output every 6 hours out to 240 h. Table 3 and Figure 39 show simple applications of tuning as well as turning on the tri-modal capability. The GF-201609 experiment simply turned on the mid-level convection and changed detrainment conditions for the momentum transport parameterization. GF may use a cloud water detrainment profile that is prescribed with c1d as a function of height. The GF-c1d experiment sets c1d to zero, and uses a cloud water detrainment profile according to mass detrainment. The introduction of c1d was shown previously to indicate the sensitivity to this parameter, especially for the temperature biases. The sensitivity is increased, depending on the microphysics parameterization, since the additional cloud water detrainment in GF-201609 causes cooling through evaporation in the low and mid-levels, while less water/ice is available in the upper levels (see Fig. 39). Large sensitivity to tuning is also shown in the wind biases.

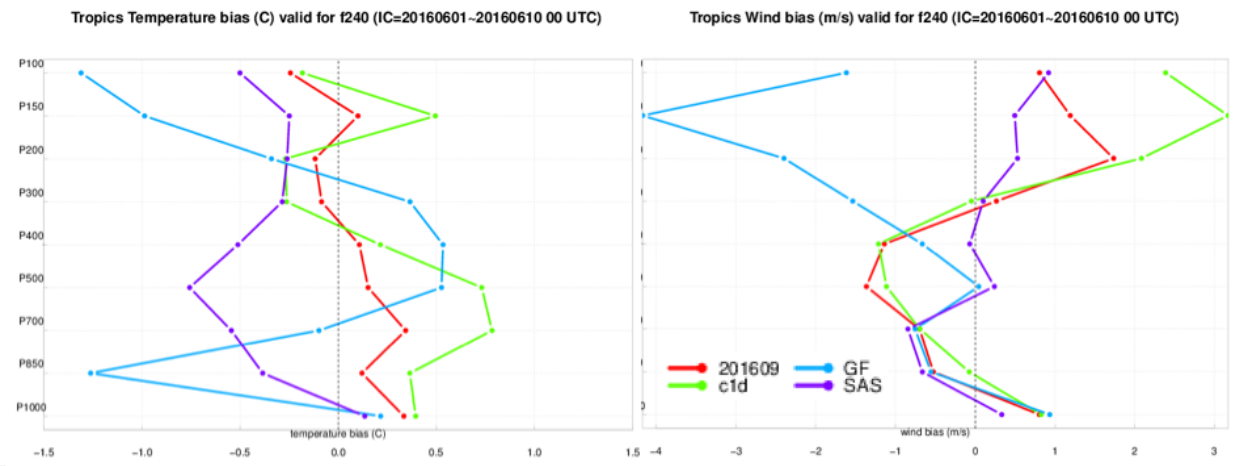
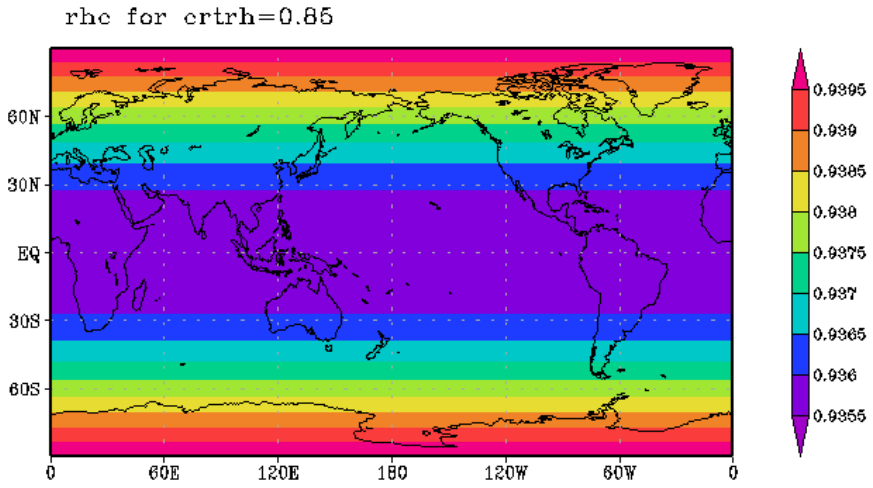


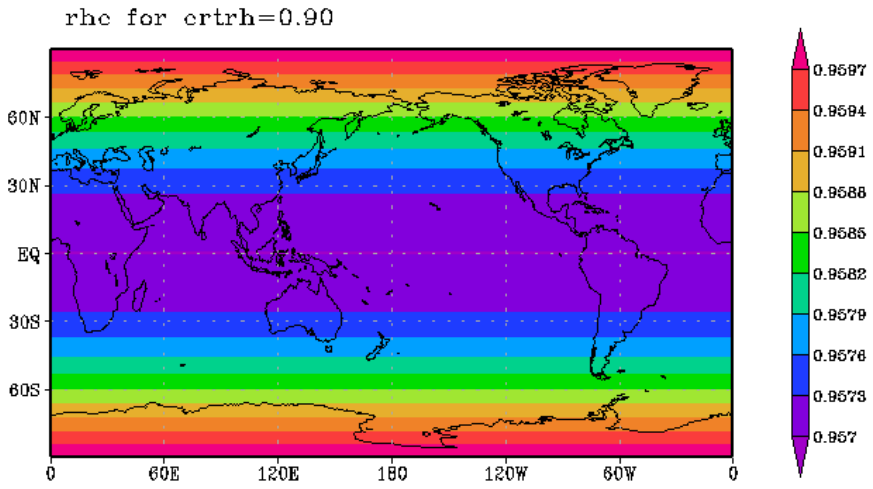
Figure 39. Vertical profile of the bias at the 240-h forecast lead time for a) temperature (°C) and b) wind (m/s) aggregated over June 1-10, 2016. GF (blue) is using the same GF convective parameterization code as used in the test by GMTB ; SAS (purple) is using the scale-aware SAS convective parameterization; GF-201609 (red) has the mid-level convection turned on as well as a modification in the momentum transport parameterization; GF-c1d (green) has the cloud water detrainment profile switched to be in line with mass detrainment.

Table 3. Convective precipitation (mm/day) and fraction compared to total precipitation averaged between 60N and 60S. Results are averaged over the 240-h forecasts from the ten forecasts from 20160601-201610. GF and SASAS using the original code as described in this document. GF-201609 and GF-c1d are described in the text.

	Convective precipitation fraction
SASAS	2.35 (65.1%)
GF	1.58 (43.2%)
GF-201609	2.80 (65.1%)
GF-c1d	3.18 (71.3%)

Tuning is not only important for a single parameterization, but instead a whole suite of parameterizations has to be taken into account. Both auto-conversion as well as a critical humidity threshold (RHC) are latitude-dependent in the GFS physics as used by the microphysics parameterization. Even though the latitudinal values are not largely different, they introduce a very significant change in the vertical biases. Figure 40 shows the latitudinal variation of RHC. Figure 41 displays vertical biases resulting from a run where RHC is set to a constant value that equals RHC in the tropics in GF-201609. As a result, we see more heating in the tropics, in spite of only a minimal difference of RHC for this area when compared to the control run (GF-201609).


rhc for crtrh=0.85
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
0
0.9995
0.999
0.9985
0.998
0.9975
0.997
0.9965
0.996
0.9955


rhc for crtrh=0.90
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
0
0.9597
0.9594
0.9591
0.9588
0.9585
0.9582
0.9579
0.9576
0.9573
0.957

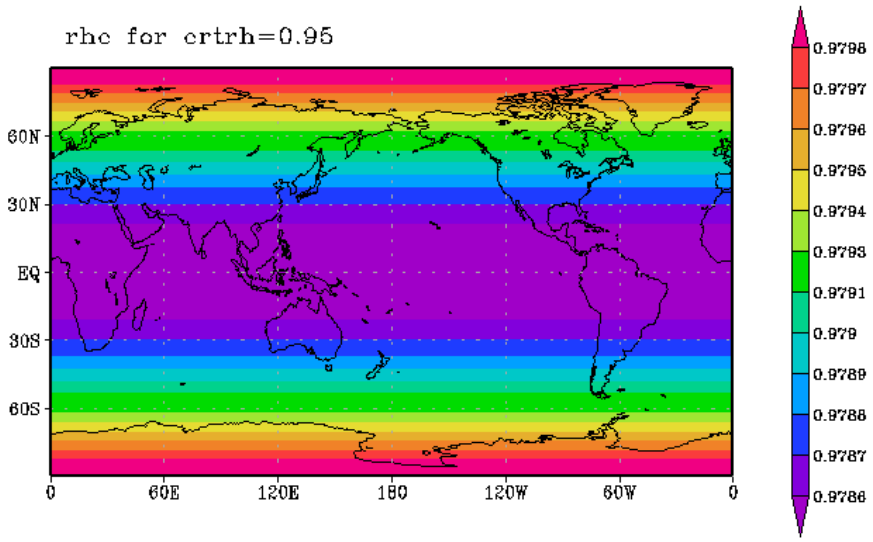

rhc for crtrh=0.95
60N
30N
EQ
30S
60S
0
60E
120E
180
120W
60W
0
0.9798
0.9797
0.9796
0.9795
0.9794
0.9793
0.9791
0.979
0.9789
0.9788
0.9787
0.9786

Figure 40. Latitudinal variation of critical relative humidity used for the microphysics parameterization.

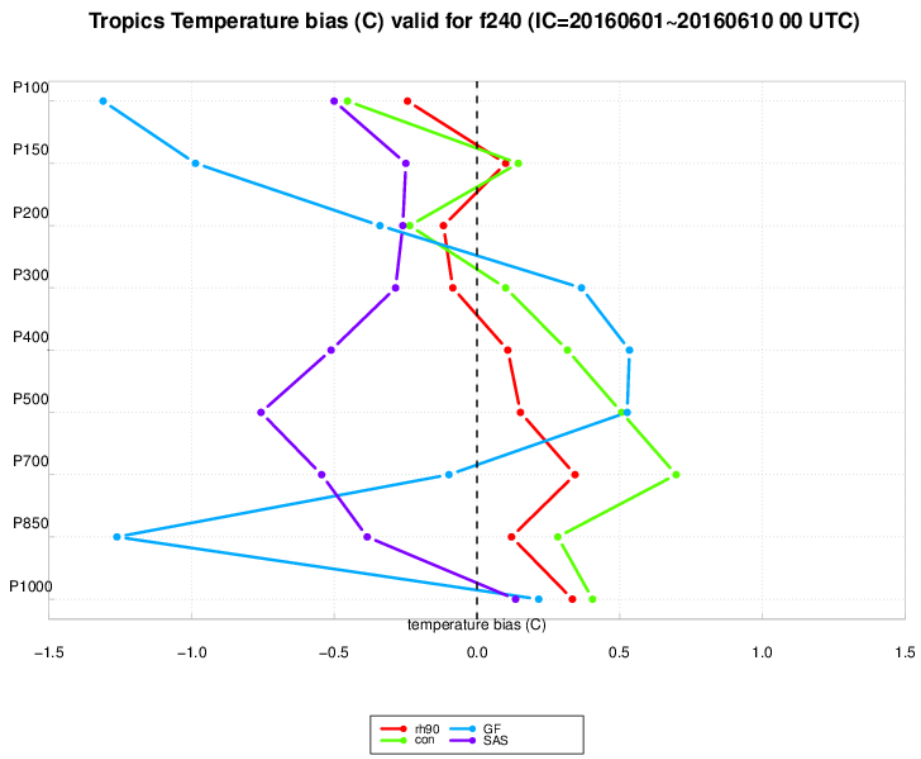


Figure 41. Vertical profile of the bias at the 240-h forecast lead time for temperature (°C) aggregated over June 1-10, 2016. GF (blue); SAS (purple); GF-201609 (red); GF with RHC set to a constant value that equals RHC in the tropics for GF-201609 (green).

Acknowledgements

The GMTB staff would like to thank the GF developer (Georg Grell of NOAA/ESRL), the NOAA/EMC Global Team led by Vijay Tallapragada, the NGGPS and EMC Strategic Implementation Plan Physics Team co-leads (Jim Doyle of NRL, Shrinivas Moorthi of NOAA/EMC, Chris Bretherton of UW, and Georg Grell), and the NGGPS Program Office led by Fred Toepfer for their engagement in preparing the test plan, defining verification metrics, supplying code, and participating in the interpretation of results. In particular, we thank EMC staff Kate Howard, for providing the GFS workflow and supporting its use, and Fanglin Yang, for exchanges regarding verification procedures. Daniel Halperin, of NCEP Weather Prediction Center, provided the code used as a starting point for the tropical cyclogenesis verification, and kindly supported GMTB in implementing it. Thanks also to Karen Griggs of NCAR for offering her desktop publishing expertise in the preparation of this document.

References

Cederwall, R.T., S.K. Krueger, S. Xie, and Yio, J, 2000: ARM/GCSS Single Column Model (SCM) Intercomparison, procedures for Case 3: Summer 1997 SCM IOP. LLNL Report, UCRL-ID-141823. Available from Atmospheric Science division, LLNL, Livermore, CA, USA.

Chikira, M. and M. Sugiyama, 2010: A cumulus parameterization with state-dependent entrainment rate. Part I: Description and sensitivity to temperature and humidity profiles. J. Atmos. Sci., **67**, 2171–2193. DOI: https://doi.org/10.1175/2010JAS3316.1.

Davies, L., C. Jakob, K. Cheung, A. D. Genio, A. Hill, T. Hume, R. J. Keane, T. Komori, V. E. Larson, Y. Lin, X. Liu, B. J., Nielsen, J. Petch, R. S. Plant, M. S. Singh, X. Shi, X. Song, W. Wang, M. A. Whitall, A. Wolf, S. Xie, and G. Zhang, 2013: A single-column model ensemble approach applied to the TWP-ICE experiment. J. Geophys. Res.: Atmos. 118, 12, 6544–6563. DOI: http://dx.doi.org/10.1002/jgrd.50450.

Grell, G.A., 1993: Prognostic Evaluation of Assumptions Used by Cumulus Parameterizations. *Mon. Wea. Rev.,* **121**, 764–787, https://doi.org/10.1175/1520-0493(1993)121<0764:PEOAUB>2.0.CO;2

Grell, G. A., and D. Dévényi, A generalized approach to parameterizing convection combining ensemble and data assimilation techniques, Geophys. Res. Lett., 29(14), doi:10.1029/2002GL015311, 2002.

Grell, G.A, and S. R. Freitas, 2014: A scale and aerosol aware stochastic convective parameterization for weather and air quality modeling. *Atmos. Chem. Phys.,* **14**, 5233- 5250. DOI: http://www.atmos-chem-phys.net/14/5233/2014/ doi:10.5194/acp-14-5233-2014.

Halperin, D., H. E. Fuelberg, R. E. Hart, and J. H. Cossuth, 2016: Verification of Tropical Cyclone Genesis Forecasts from Global Numerical Models: Comparisons between the North Atlantic and Eastern North Pacific Basins. *Wea. and Forec., **31**, 947-955.* DOI: http://journals.ametsoc.org/doi/10.1175/WAF-D-15-0157.1.

Han, J., W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017. Updates in the NCEP GFS Cumulus Convection Schemes with Scale and Aerosol Awareness. Submitted to *Wea. and Forec.*

Huffman, G. J., D.T. Bolvin, and E. J. Nelkin, 2010: The TRMM Multi-Satellite Precipitation Analysis (TMPA). *Satellite Rainfall Applications for Surface Hydrology*, M.Gebremichael and F. Hossain, Eds., Springer, 3-22.

Randall, D. A., and D. G. Cripe, 1999: Alternative methods for specification of observed forcing in single-column models and cloud system models. J. Geophys. Res.: Atmos., 104, D20, 24527–24545. DOI: http://dx.doi.org/10.1029/1999JD900765.

Randall, D., S. Krueger, C. Bretherton, J. Curry, P. Duynkerke, M. Moncrieff, B. Ryan, D. Starr, M. Miller, W. Rossow, G. Tselioudis, and B. Wielicki, 2003: Confronting models with data - the GEWEX cloud systems study. Bull. Amer. Meteor. Soc., 84, 455–469. DOI: http://dx.doi.org/10.1175/BAMS-84-4-455.

Taylor, K., 2001: Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmos. 106, D7, 7183–7192. DOI: http://dx.doi.org/10.1029/2000JD900719

Xie, S., Xu, K.-M., Cederwall, R. T., Bechtold, P., Genio, A. D. D., Klein, S. A., Cripe, D. G., Ghan, S. J., Gregory, D., Iacobellis, S. F., Krueger, S. K., Lohmann, U., Petch, J. C., Randall, D. A., Rotstayn, L. D., Somerville, R. C. J., Sud, Y. C., Salzen, K. V., Walker, G. K., Wolf, A., Yio, J. J., Zhang, G. J., and Zhang, M. Intercomparison and evaluation of cumulus parameterizations under summertime midlatitude continental conditions. Quarterly Journal of the Royal Meteorological Society 128, 582 (2002), 1095–1135.

Zhang, M., R. C. J. Somerville, and S. Xie, 2016: The SCM concept and creation of ARM forcing datasets. Meteor. Monogr. 57, 24.1–24.12. DOI: http://dx.doi.org/10.1175/AMSMONOGRAPHS-D-15-0040.1.

Appendix A. List of acronyms

AC: Anomaly Correlation
AL: Atlantic TC basin
ARM: Atmospheric Radiation Measurement
BUFR: Binary Universal Form for Representation of Meteorological Data
CI: Confidence Interval
CCPA: Climatology-Calibrated Precipitation Analysis
CCPP: Common Community Physics Package
CONUS: Contiguous United States
CMORPH: Climate Prediction Center Morphing Technique
DA: Data Assimilation
DTC: Developmental Testbed Center
EMC: Environmental Modeling Center
ETS: Equitable Threat Score
ESRL: Earth System Research Laboratory
EP: Eastern North Pacific TC basin
FBias: Frequency Bias
GCSS: GEWEX Cloud System Study
GD_KF: Key Finding from diagnostics of global single case
GDAS: Global Data Assimilation System
GEWEX: Global Energy and Water cycle EXchanges
GF: Grell-Freitas cumulus parameterization
GFS: Global Forecast System
GFS-GF: Global Forecast System run with GF in cycled DA mode
GFS-GFcold: Global Forecast System run with GF in cold start mode
GFS-SAS: Global Forecast System run with scale aware SAS in cycled DA mode
GISS: NASA Goddard Institute for Space Studies
GMTB: Global Model Test Bed
GRIB2: GRIdded Binary file format version2
GSD: Global Systems Division
GSM: Global Spectral Model
GV_KF: Key Finding from analysis of verification of global retrospective runs
HPSS: High Performance Storage System
HWRF: Hurricane Weather Research and Forecasting System
ITCZ: Intertropical Convergence Zone
JTWC: Joint Typhoon Warning Center
LES: Large Eddy Simulation
MET: Model Evaluation Tools
METAR: aviation routine weather report
NAM: North American Mesoscale Forecast System
NCAR: National Center for Atmospheric Research
NCEP: National Centers for Environmental Prediction
NDAS: NAM Data Assimilation System
NEMS: NOAA Environmental Modeling System
NEMSIO: NEMS Input/Output format
NGGPS: Next-Generation Global Prediction System

NH: Northern Hemisphere (defined here as 20^o – 80^o N for upper air verification and 20^o – 60^o N for precipitation verification)
NHC: National Hurricane Center
NOAA: National Oceanic and Atmospheric Administration
PBL: Planetary Boundary Layer
PrepBUFR: Quality-controlled BUFR
RAOB: RAwinsonde OBservation
RAP: Rapid Refresh Forecast System
RMSE: Root-Mean-Square Error
RRTM: Rapid Radiative Transfer Model
RRTMG: RRTM for General Circulation Models
R2O: Transition of Research to Operations
SAS: Simplified Arakawa-Schubert cumulus parameterization
SCM: Single Column Model
SCM_KF: Key Finding from analysis of Single Column Model results
SH: Southern Hemisphere (defined here as 20^o – 80^o S for upper air verification and 20^o – 60^o S for precipitation verification)
SPCZ: South Pacific Convergence Zone
SS**:** Statistically Significant
SST: Sea Surface Temperature
SVN: Apache Subversion
TMPA: TRMM Multisatellite Precipitation Analysis
TRMM: Tropical Rainfall Measuring Mission
TC: Tropical Cyclone
TG: Tropical Cyclogenesis
TROP: Tropics (defined here as 20^o S – 20^o N)
TWP-ICE: Tropical Warm Pool - International Cloud Experiment
UPP: Unified Post Processor
UTC: Coordinated Universal Time
VLab: Virtual Laboratory
WP: West Pacific TC basin