Quality Control in CICE: Why is it Necessary and How Can it be Performed?

Andrew Roberts[1]
Rick Allard[2]
[1]Los Alamos National Laboratory
[2]Naval Research Laboratory

Overview

- New contributions (e.g., physics, biogeochemistry) to CICE Consortium code should not change the physics of existing model configurations when switched off.
- CICE must reproduce answers bit-for-bit (bfb) as compared to previous simulations with the same namelist configurations.
- However, some model changes (e.g., bug fixes, new physics etc.) may not produce bfb results; thus requiring bfb testing to confirm or deny the null hypothesis, which is that new additions to the CICE dynamical core and CICE have not significantly altered simulated sea ice volume using previous model configurations.
- Here we present a methodology to perform quality control (QC) testing with CICE.

The Theoretical Basis for the CICE Quality Control Algorithm

PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Cite this article: Roberts AF, Hunke EC, Allard R, Bailey DA, Craig AP, Lemieux J-F, Turner MD. 2018 Quality control for community-based sea-ice model development. *Phil. Trans. R. Soc. A* **376**: 2017.0344. http://dx.doi.org/10.1098/rsta.2017.0344

Accepted: 16 July 2018

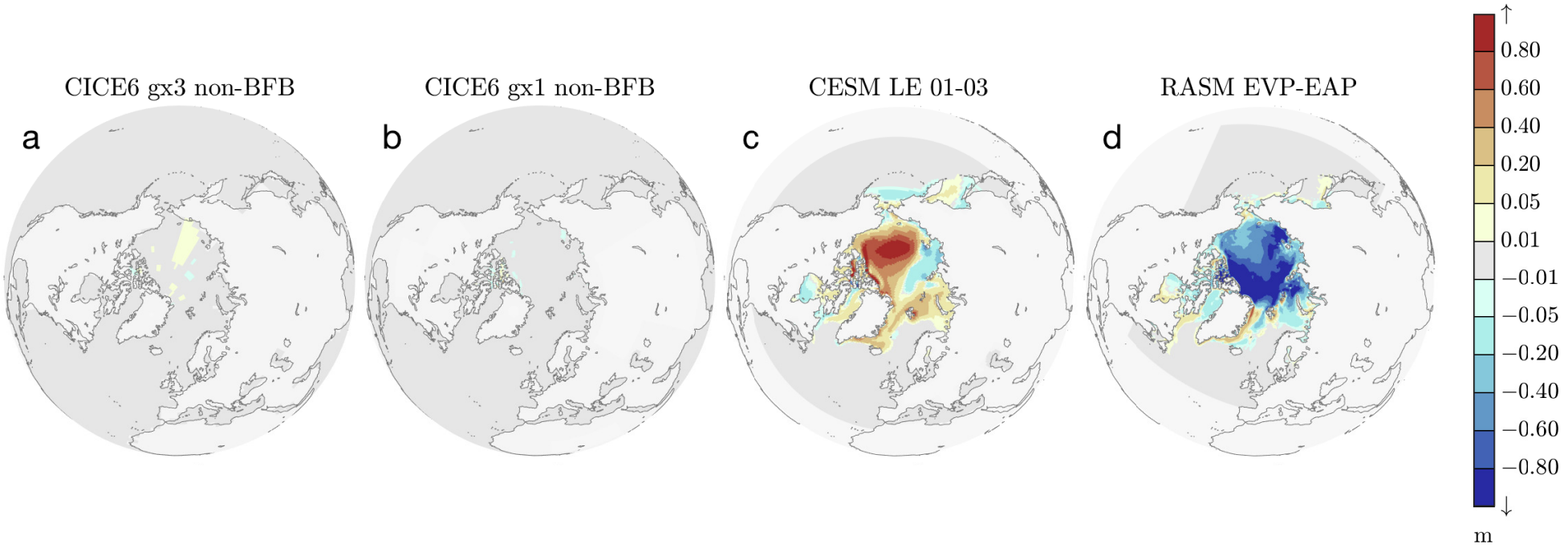
One contribution of 15 to a theme issue 'Modelling of sea-ice phenomena'.

Quality control for community-based sea-ice model development

Andrew F. Roberts[1], Elizabeth C. Hunke[2], Richard Allard[3], David A. Bailey[4], Anthony P. Craig[5], Jean-François Lemieux[6] and Matthew D. Turner[7]

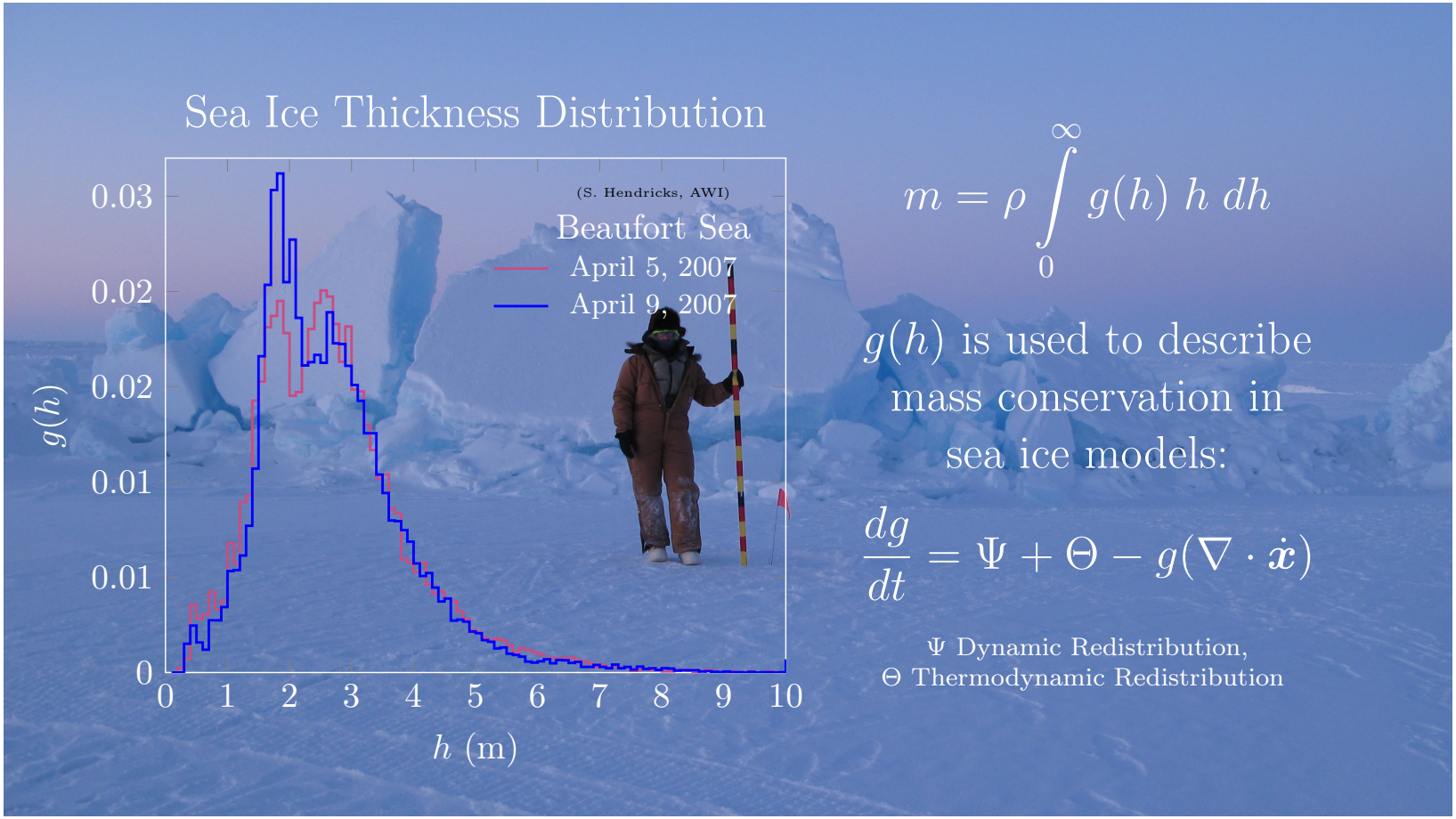
[1]Naval Postgraduate School, Monterey, CA, USA
[2]Los Alamos National Laboratory, Los Alamos, NM, USA
[3]Naval Research Laboratory, Stennis Space Center, MS, USA
[4]National Center for Atmospheric Research, Boulder, CO, USA
[5]Cherokee Nation Technologies in support of NOAA Earth System Research Laboratory, Washington, DC, USA
[6]Recherche en Prévision Numérique Environnementale, Environnement et Changement Climatique Canada Dorval, QC, Canada
[7]DoD HPCMP PETTT, Engility Corp., Stennis Space Center, MS, USA

What change between model simulations constitutes a change in sea ice climate for the purpose of quality control?



Above: Examples of ice thickness differences between small changes in CICE (a,b), between CESM-LE ensemble members, (c), and between EAP and EVP in RASM (d).

How do we define a change in climate?



Judging quality control in models is a developing science.
We have chosen to use the core state variable in CICE: $g(h)$

How do we categorize the quality of a code version against an established baseline?

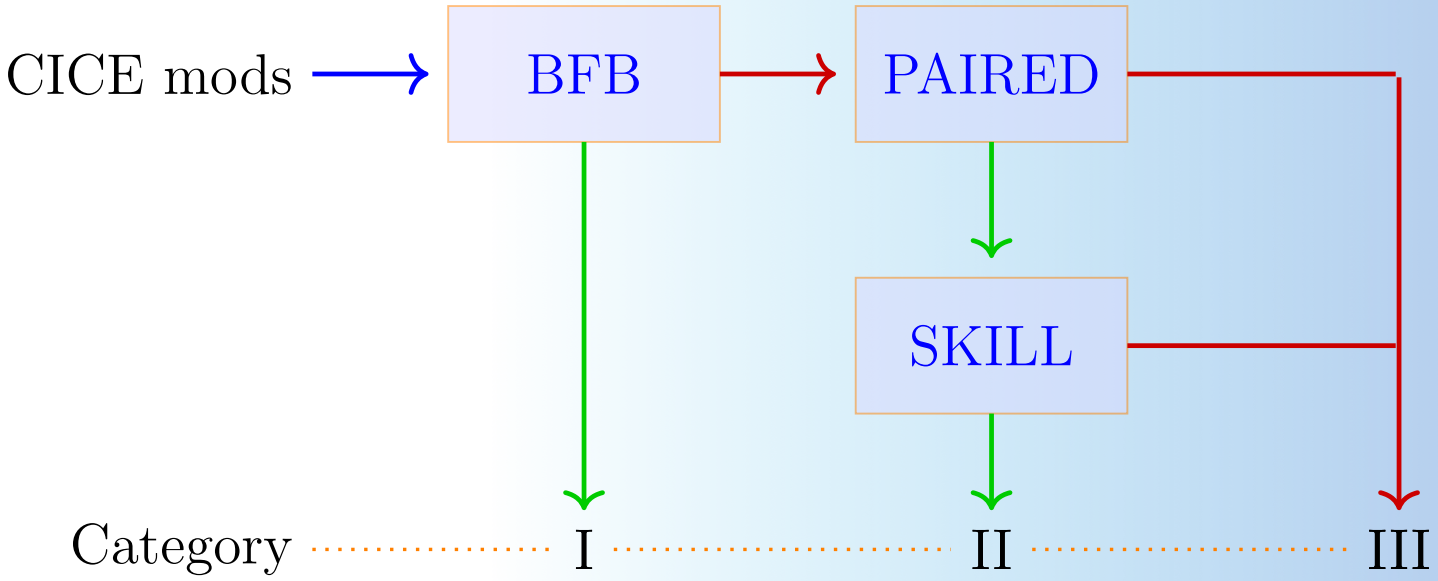
Category I: Bit-for-bit (BFB) change to the code

Category II: Not BFB, but not climate changing

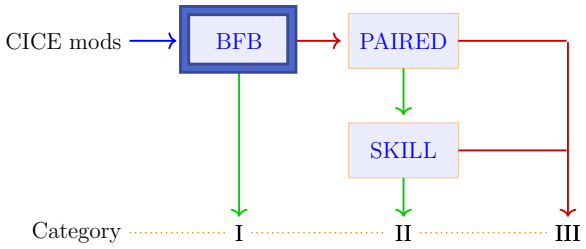
Category III: Climate changing

Categort IV: New or corrected physics subject to scientific review

How do we define a change in sea ice climate?

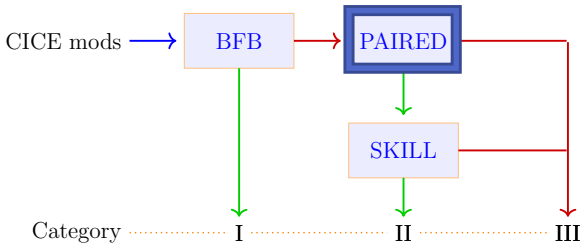


If a code change does not pass a BFB test, we determine a change in climate using a paired thickness test and a sea ice model skill test.



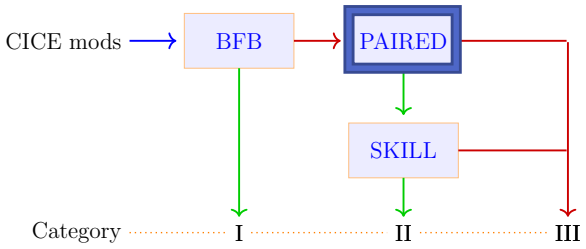
The BFB test

Simply compare log files representing the same time period between the new simulation and the baseline, using tools such as diff or vimdiff.



The paired thickness test

If the BFB test fails, we now have to determine statistically if the baseline climate is the same as the climate of the changed-code climate. The simplest way to do this is with the two-stage paired thickness test.



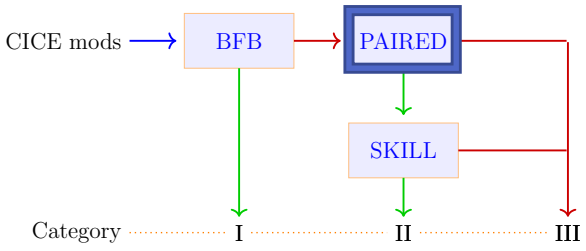
The paired thickness test

***Stage 1**:* For all locations on the CICE gx1 model domain where ice thickness is greater than 0.01m (we define this as the sea-ice zone for our purpose), determine whether the null hypothesis is true at the 80% confidence interval using:

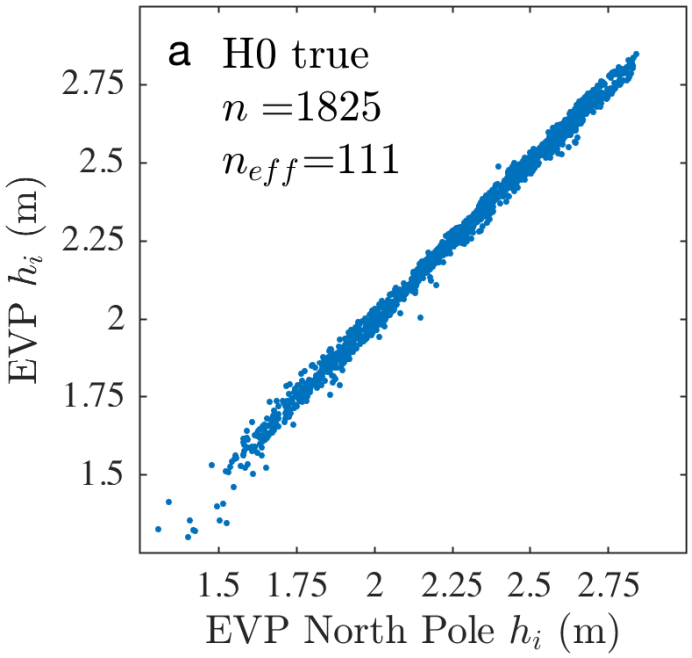
$$t = \frac{\bar{h}_\Delta}{\sigma_\Delta / \sqrt{n_{eff}}}$$

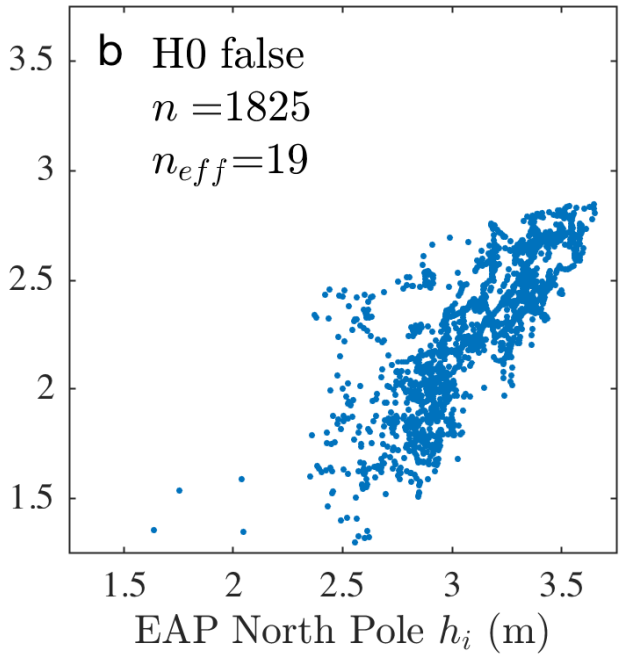
Where $n_{eff} = n(1 - r_1)/(1 + r_1)$ and r_1 is the lag-1 autocorrelation. If n_{eff}<30, then the test becomes conservative, meaning that it can erroneously indicate no difference between a simulation with changed code and the baseline.

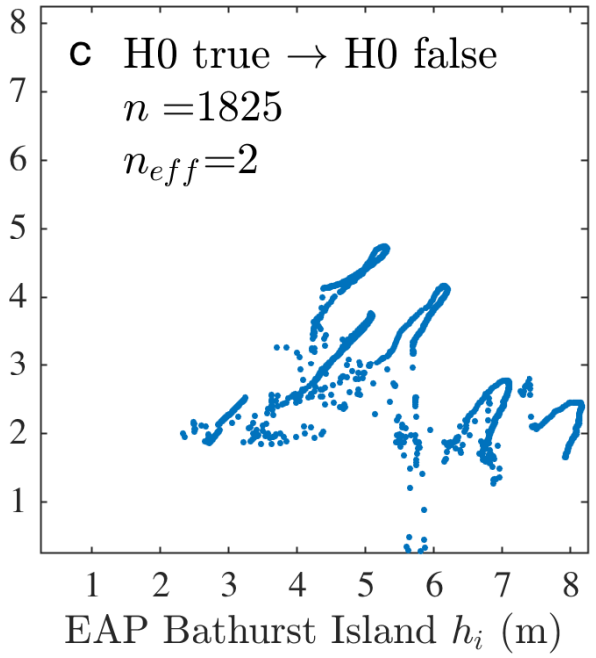
Here $\bar{h}_\Delta$ is the grid-cell difference in mean sea ice thickness, and σ_Δ is the paired grid-cell thickness difference standard deviation of the series $h_\Delta = h_{a_i}$- h_{b_i}.



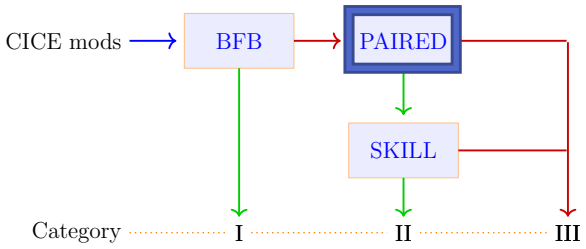
The paired thickness test







Above: Example of an erroneous confirmation of non-climate changing code in (c)



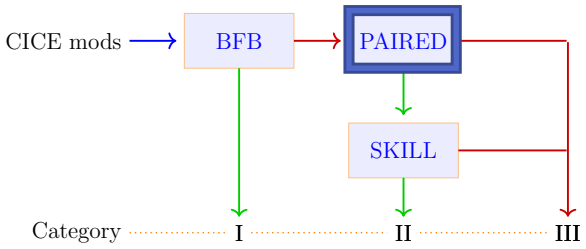
The paired thickness test

***Stage 1**:* For all locations on the CICE gx1 model domain where ice thickness is greater than 0.01m (we define this as the sea-ice zone for our purpose), determine whether the null hypothesis is true at the 80% confidence interval using:

$$t = \frac{\bar{h}_\Delta}{\sigma_\Delta / \sqrt{n_{eff}}}$$

Where $n_{eff} = n(1 - r_1)\ (1 + r_1)$ and r_1 is the lag-1 autocorrelation. If n_{eff}<30, then the test becomes conservative, meaning that it can erroneously indicate no difference between a simulation with changed code and the baseline.

If n_{eff} ≥30, the answer stands, and we use the outcome to indicate if a code-change is climate-changing. If not, we progress to ***Stage 2.***



The paired thickness test

Stage 2. If n_{eff}<30, and the null hypothesis is confirmed, we now defer to a lookup table using the standard t-test generated with a Monte-Carlo methods:

$$t = \frac{\bar{h}_\Delta}{\sigma_\Delta / \sqrt{n}}$$

where n appears in the equation instead of n_{eff}.

Lookup table for a 5-year simulation.

Table 2. Critical *t*-values for Stage 2 of the Two-Stage Paired Thickness Test (2SPT) generated from 10 million AR(1) timeseries of length $n = 1825$ ($N = 1824$) for lag-1 autocorrelation r_1 and two-sided tests at the 80% and 95% confidence intervals using the method described in the appendix. The length of the AR(1) series used here corresponds to a 5-year sequence of daily ice thickness model archives using a no-leap proleptic Gregorian calendar frequently employed in sea-ice models, but values change little by increasing the sample size to $n = 1827$ to accommodate two leap days possible within a 5-year series. Values at $r_1 = 0$ (blue) represent the standard critical *t*-statistic for uncorrelated samples.

r_1	−0.10	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80
80%	1.18	1.28	1.42	1.57	1.76	1.97	2.23	2.59	3.05	3.88
95%	1.80	1.96	2.17	2.43	2.67	3.01	3.44	3.98	4.72	5.99
r_1	0.82	0.84	0.86	0.88	0.90	0.91	0.92	0.93	0.94	0.95
80%	4.12	4.38	4.70	5.15	5.64	6.03	6.41	6.95	7.57	8.35
95%	6.36	6.78	7.30	8.00	8.80	9.33	10.10	10.72	11.81	13.14
r_1	0.96	0.97	0.98	0.99	0.992	0.994	0.996	0.998	0.999	
80%	9.44	11.07	14.29	23.01	27.03	33.05	40.76	49.52	53.94	
95%	14.89	18.16	23.88	43.22	52.29	62.89	73.10	81.69	84.91	

The lookup table is most sensitive at high r_1 values and small n.

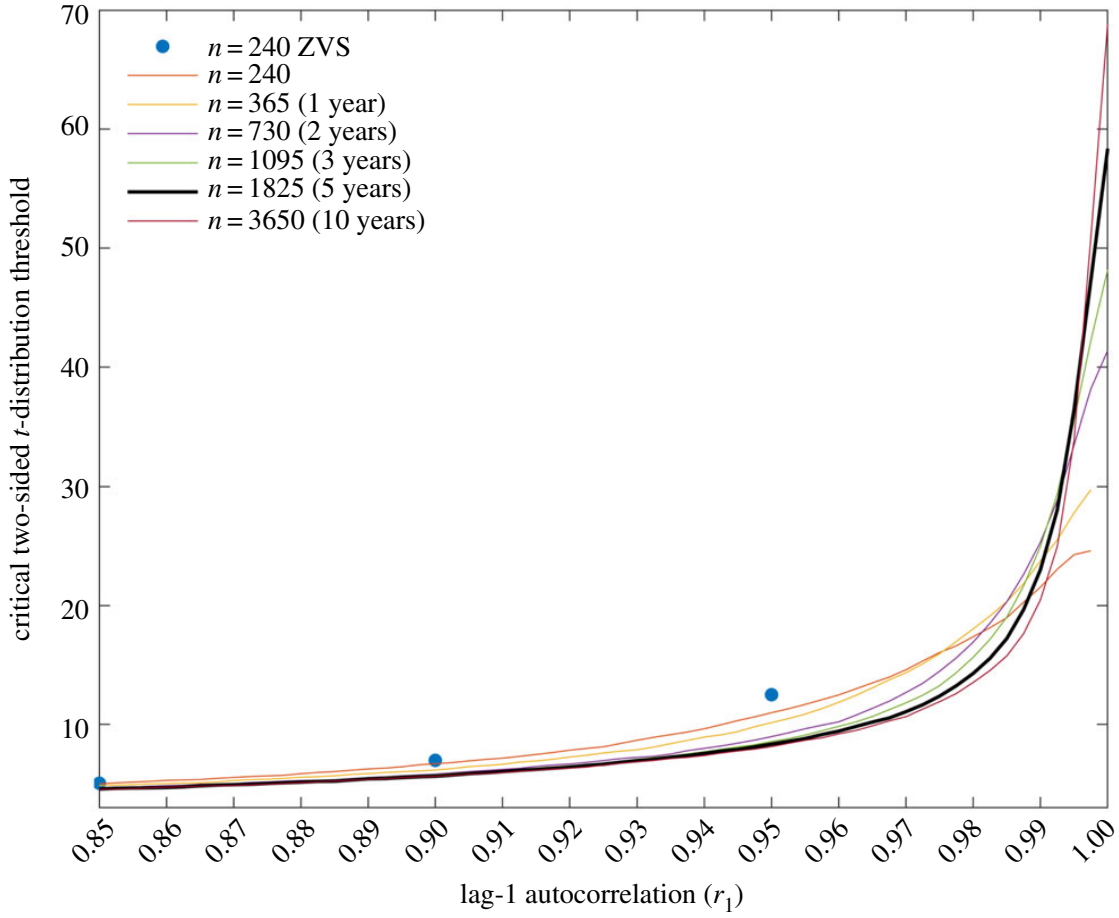
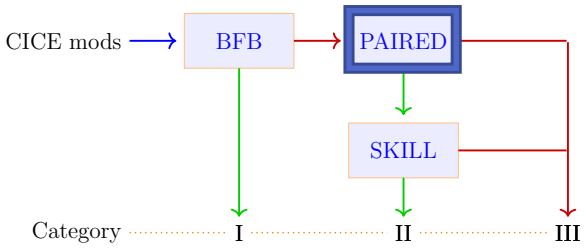
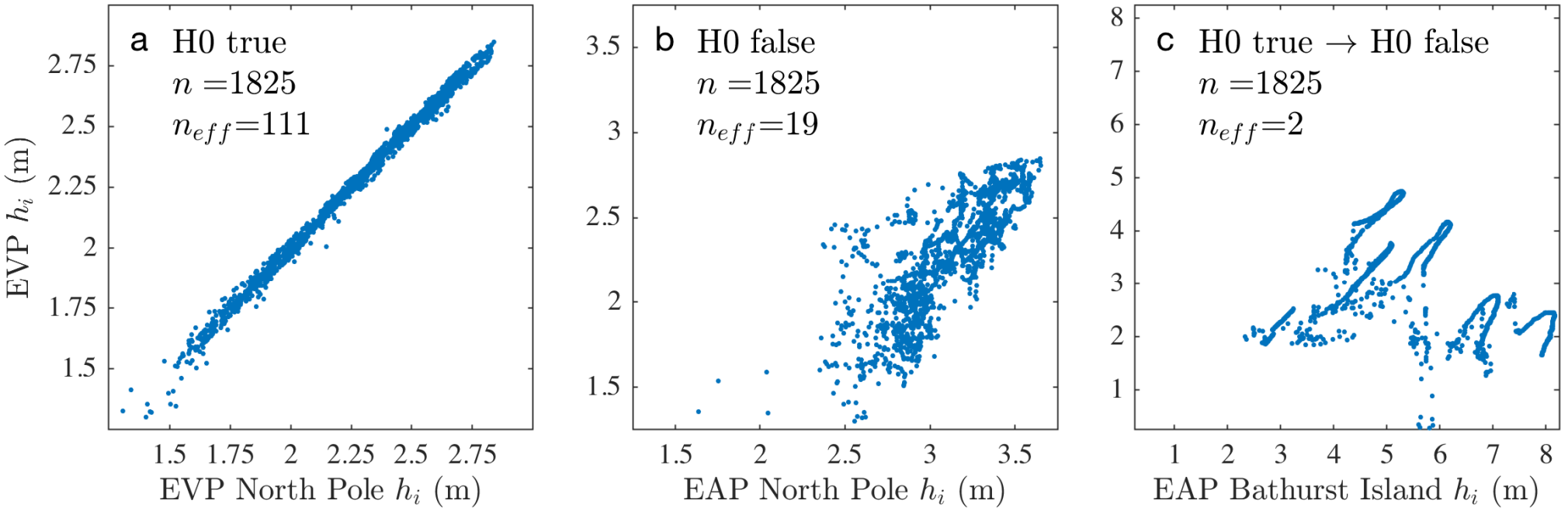


Figure 7. Critical *t*-statistics at the high end of the r_1 scale for the 80% two-sided confidence interval generated using the method described in the appendix. Change in the statistic with increasing sample sizes is indicated for the maximum sample size explored by Zwiers and von Storch (ZVS) in [39], $n = 240$, out to the equivalent of a 10-year series of daily thickness samples from sea ice models, $n = 3650$ (no-leap calendar). Tabulated values from Zwiers & von Storch [39] appear as blue data points and are comparable with the $n = 240$ red trace generated using the large ensemble method used in this paper. The statistic for the baseline series length used by the CICE Consortium, $n = 1825$, appears in bold black.

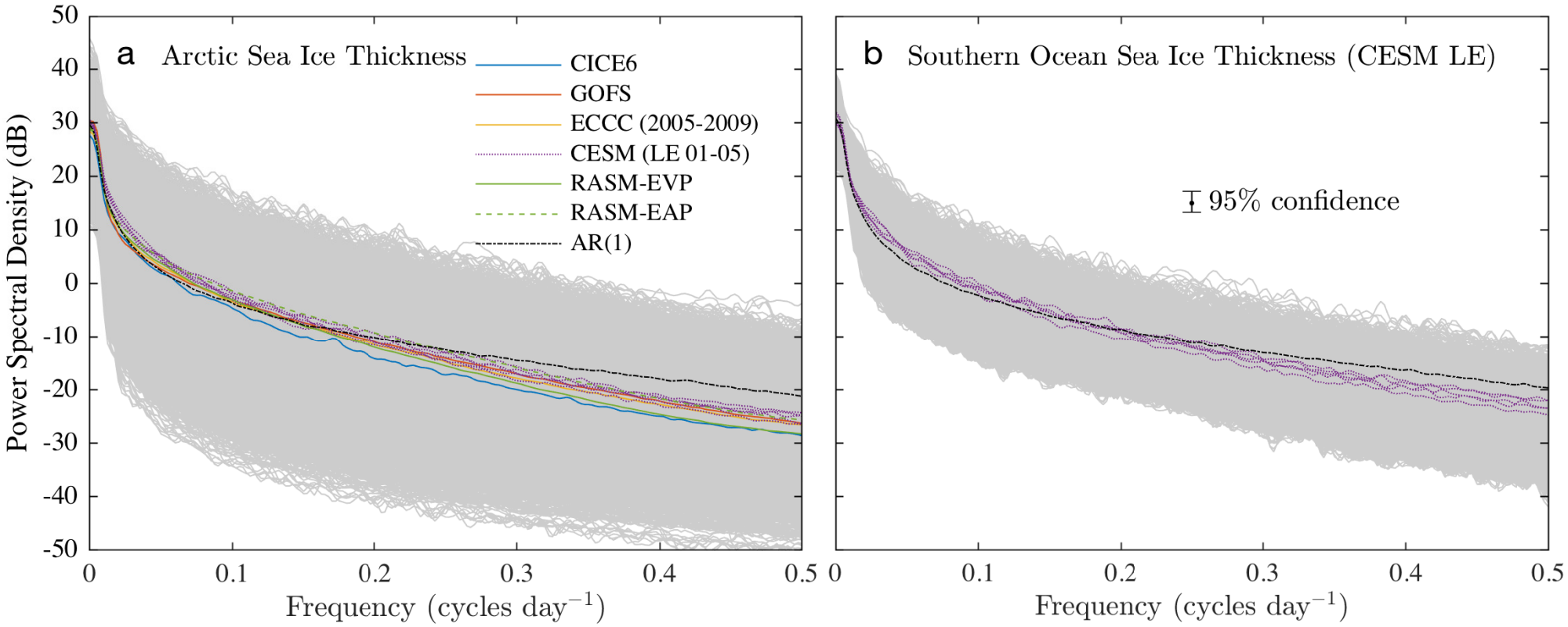


The paired thickness test



Above: Example of an erroneous confirmation of non-climate changing code in (c)

This statistical test can only be justified if sea ice thickness evolution can be well approximated by an AR(1) process, which it can, as shown here for CICE Consortium models:



The AR(1) model here is given by $h_i = 0.994\ h_{i-1} + \varepsilon_i$ for thickness timeseries h_i and white noise ε_i.

Each of these models are very different, but possess nearly identical statistical properties in the evolution of thickness within a grid cell.

model[a]	lead[b]	configuration[c]	domain	CICE[d]	thermodynamics [12,19]	radiation [17,18]	melt ponds [14,15]	dynamics[e]
CICE6 [5,6]	LANL	ice	global	6.0	Mushy Layer	Delta-Eddington	Level Ice	EVP
GOFS [2]	NRL	ocn-ice-assim	global	4.0	Bitz–Lipscomb	CCSM3	—	EVP
ECCC [3,20]	ECCC	ocn-ice	regional	4.0	Bitz–Lipscomb	CCSM3	—	EVP/landfast Ice
RASM [21,22]	NPS	ocn-ice-atm-Ind	regional	5.1	Mushy Layer	Delta-Eddington	Level Ice	EVP and EAP
CESM [23]	NCAR	ocn-ice-atm-Ind	global	4.1	Bitz–Lipscomb	Delta-Eddington	CESM	EVP

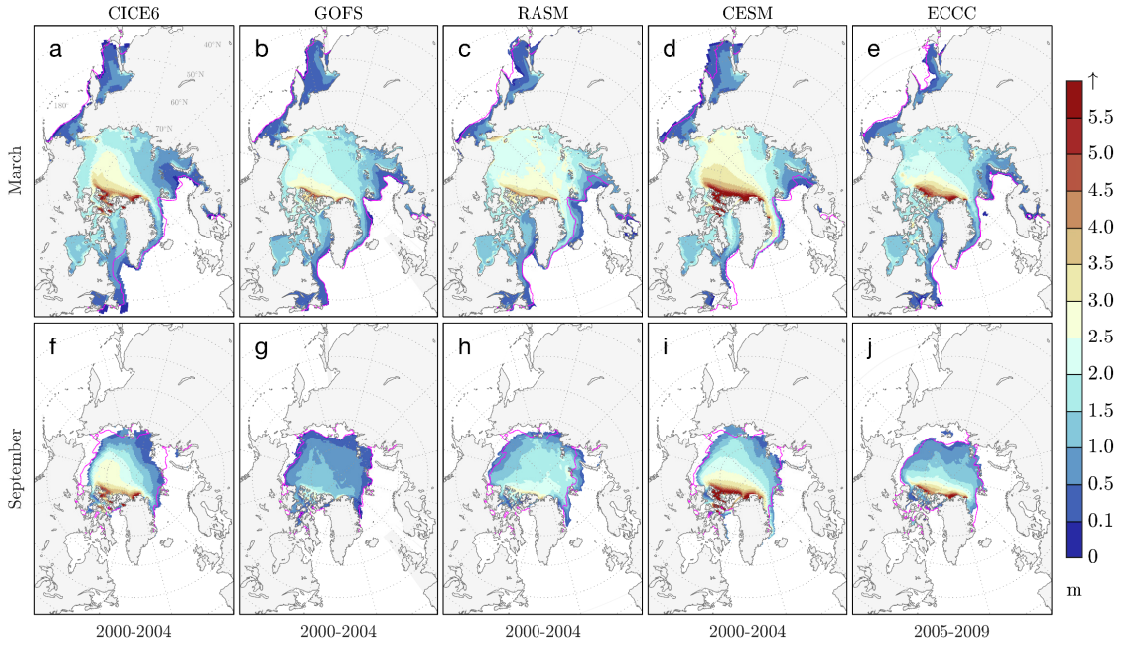
[a]CICE6, CICE Consortium dynamic core with Icepack; GOFS, US Navy Global Ocean Forecast System v. 3.1; ECCC, Environment and Climate Change Canada model; RASM, Regional Arctic System Model v. 1.1; CESM, Community Earth System Model Large Ensemble.

[b]LANL, Los Alamos National Laboratory; NRL, Naval Research Laboratory; ECCC, Environment and Climate Change Canada; NPS, Naval Postgraduate School; NCAR, National Center for Atmospheric Research.

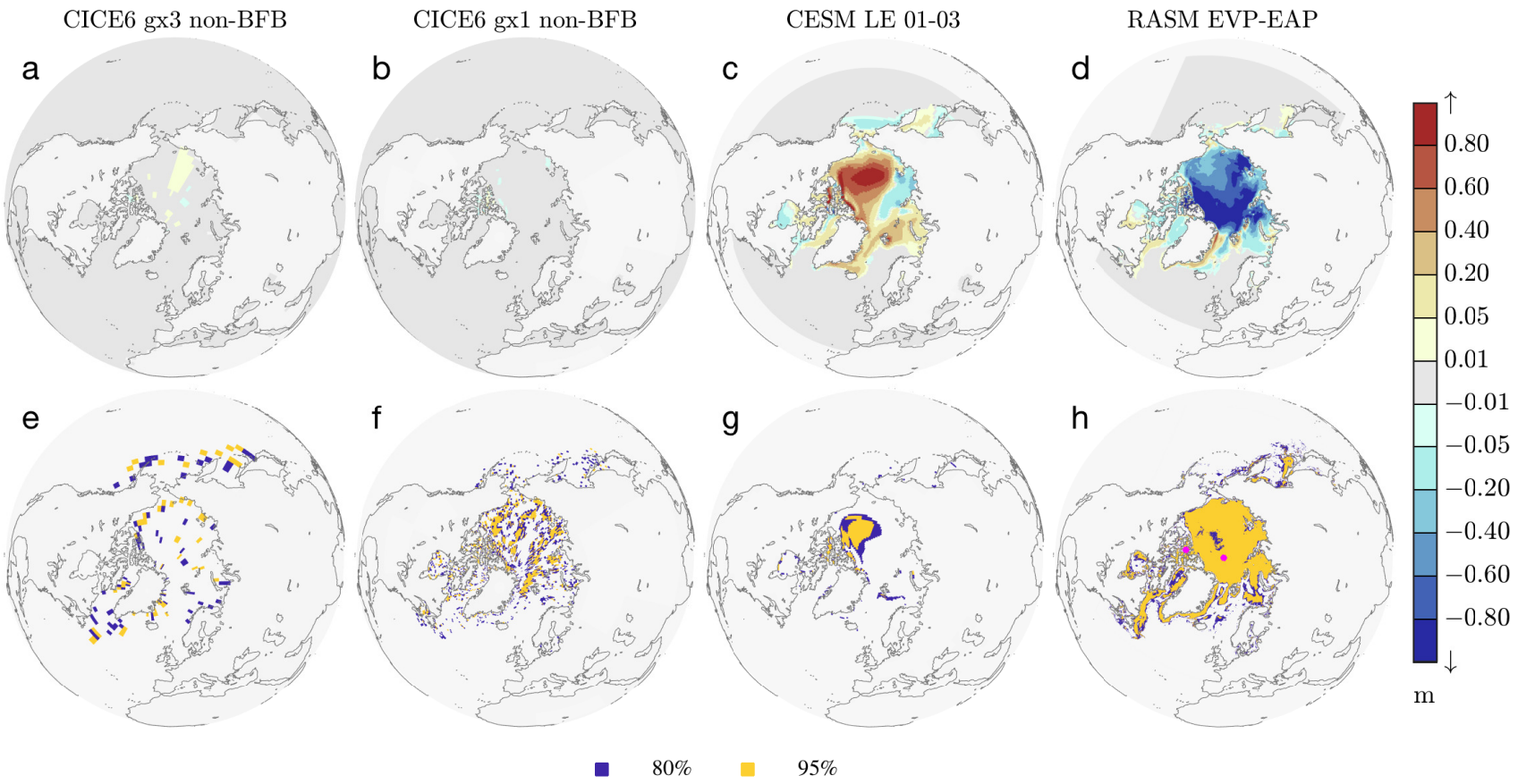
[c]ice - standalone sea ice model; ocn-ice - coupled ocean and ice model forced with atmospheric reanalyses; ocn-ice-assim - assimilated and coupled ocean and ice model forced with atmospheric reanalyses; ocn-ice-atm-Ind - fully coupled ocean, sea ice, atmosphere and terrestrial models, forced laterally with observation-based datasets if regional, or with transient greenhouse gas concentrations if global.

[d]CICE code v. 4 [24], 5 [4] or 6 [5,6].

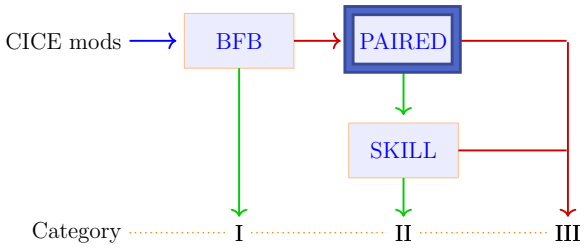
[e]EVP, elastic-viscous-plastic [7,8]; EAP, elastic-anisotropic-plastic [10].



What change in a model's code constitutes a change in sea ice climate?



A minimum of 50% of the sea ice zone must fail the test to be a Category III



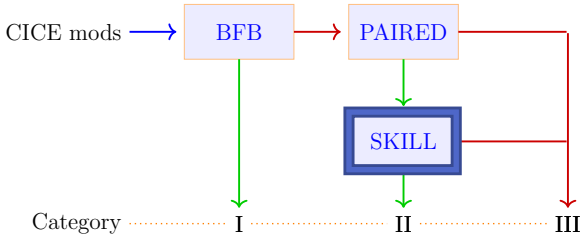
The paired thickness test

Stage 2. If n_{eff}<30, and the null hypothesis is confirmed, we now defer to a lookup table using the standard t-test generated with a Monte-Carlo methods:

$$t = \frac{\bar{h}_\Delta}{\sigma_\Delta / \sqrt{n}}$$

where n_{eff} appears in the equation instead of n.

Categorization Stage. Calculate the area-weighted fraction of the test regions that failed (i.e. where H1 is true). If the outcome is less than 50% of the sea ice zone with the alternate hypothesis confirmed, the test passes as Category II and proceeds to our final test, otherwise our QC algorithm stops and is labeled Category III.

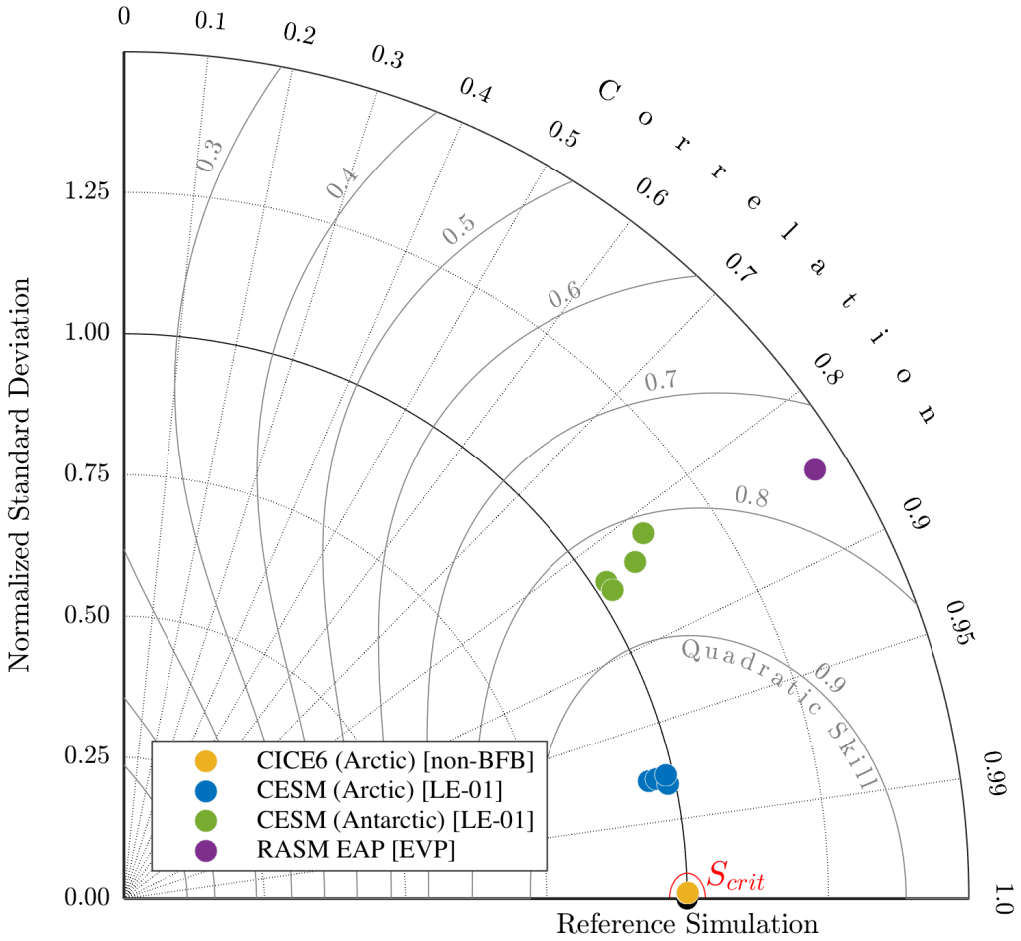


The final skill test

The skill test constitutes a test of the pan-Arctic and pan-Antarctic thickness patterns, using the Quadratic Skill score:

$$S = \left[\frac{(1+R)\sigma_a\sigma_b}{(\sigma_a^2+\sigma_b^2)}\right]$$

where R is the correlation coefficient between two vectors of co-located thicknesses from two simulations, and σ_a and σ_b are the standard deviations of the respective vectors. Values within the red circle (right) pass.



Generating Quality Control Test Cases

On cheyenne (or your own machine):

cp ./CICE/configuration/scripts/tests/QC/gen_qc_cases.csh to ~/CICE

To load the Python module and activate the virtual environment to access numpy and matplotlib, type:

$ module load python

$ ncar_pylib

Now run the script:

$ ./gen_qc_cases.csh --machine cheyenne (default is for gx3 grid; use "-g gx1" for gx1 tests)

It will generate the following test cases and directories:

1) Base case: **./CICE/cheyenne_intel_smoke_gx3_4x1_long_qc.qc_base**
2) BFB case: **./CICE/cheyenne_intel_smoke_gx3_4x1_long_qc.qc_bfb**
3) Non-BFB but not climate changing:
   ./CICE/cheyenne_intel_smoke_gx3_4x1_long_qc_nonbfb.qc_test
4) Non-BFB and climate changing:
   ./CICE/cheyenne_intel_smoke_gx3_4x1_alt02_long_qc.qc_fail

Differences Between the 4 CICE Test Cases

- All use gx3 global grid (3°x3°), NCAR bulk atmospheric forcing, and are run for 5 years
- The base (Case 1) and bit-for-bit [BFB] (Case 2) tests are identical
- Case 3 (non-BFB but not climate changing) differs from base case with a timestep of 1800 versus 3600, with increased iteration count = 87,600)

Ice initialization

Restart file switch

# of ice categories

Melt ponds on/off

Shortwave formulation

	Ice_ic	restart	kcatbound	ncat	tr_pond_lvl	kitd	revised_evp	kstrength	shortwave	distrubution type
Case 3	default'	.false.	-1	1	.false.	0	.true.	0	'cssm3'	sectrobin
Case 4	../cice_consortium/CICE_data/ic/gx3	.true.	0	5	.true.	1	.false.	1	dEdd'	cartesian

Significant differences between Case 3 and "climate changing" Case 4

Performing QC Analysis with CICE

- Test cases are run (only take a few hours for gx3 grid) and generate daily output (history) files for a 5-year model run with gx3 test case.
- The script below invokes ./configuration/scripts/tests/QC/cice.t-test.py to perform t-test validation for non-bit-for-bit results for CICE
- $ ./compare_qc_cases.csh

Running QC test on the following directories:

/glade/scratch/rallard/CICE_RUNS/cheyenne_intel_smoke_gx3_4x1_long_qc.qc_base/history

/glade/scratch/rallard/CICE_RUNS/cheyenne_intel_smoke_gx3_4x1_long_qc.qc_bfb/history

Number of files: 1825 (365 days X 5 years)

Data is bit-for-bit. No need to run QC test

Performing QC Analysis with CICE (cont.)
Two-Stage Paired Thickness Test and Quadratic Skill Compliance Test

/glade/scratch/rallard/CICE_RUNS/gordon_intel_smoke_gx3_4x1_long_qc.qc_base/history

/glade/scratch/rallard/CICE_RUNS/gordon_intel_smoke_gx3_4x1_long_qc_nonbfb.qc_test/history

Number of files: 1825

2-Stage Test Passed (confirms that the 2 simulations paired differences are ~0)

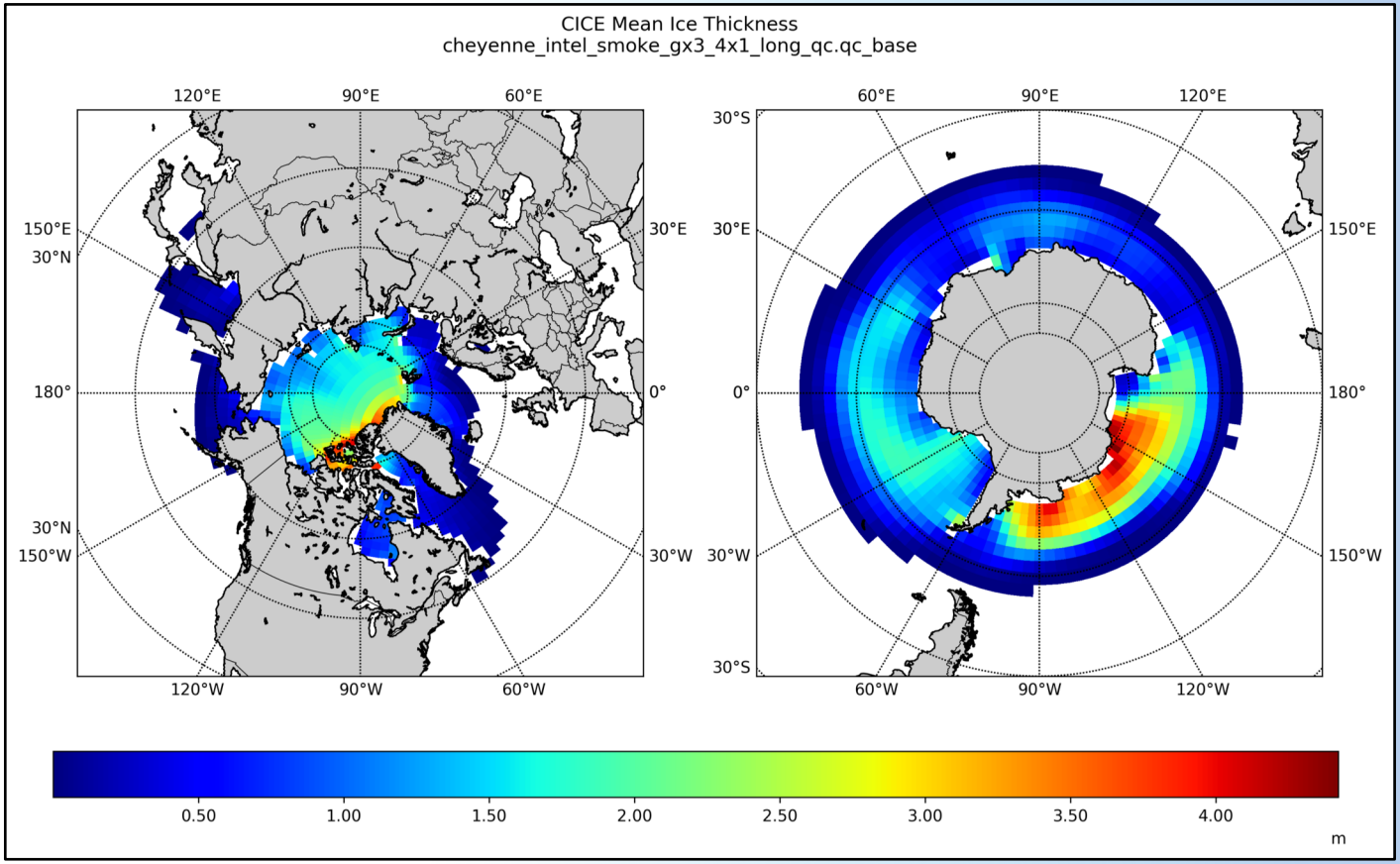
Quadratic Skill Test Passed for Northern Hemisphere ($S_{crit} \geq 0.99$)

Quadratic Skill Test Passed for Southern Hemisphere ($S_{crit} \geq 0.99$)

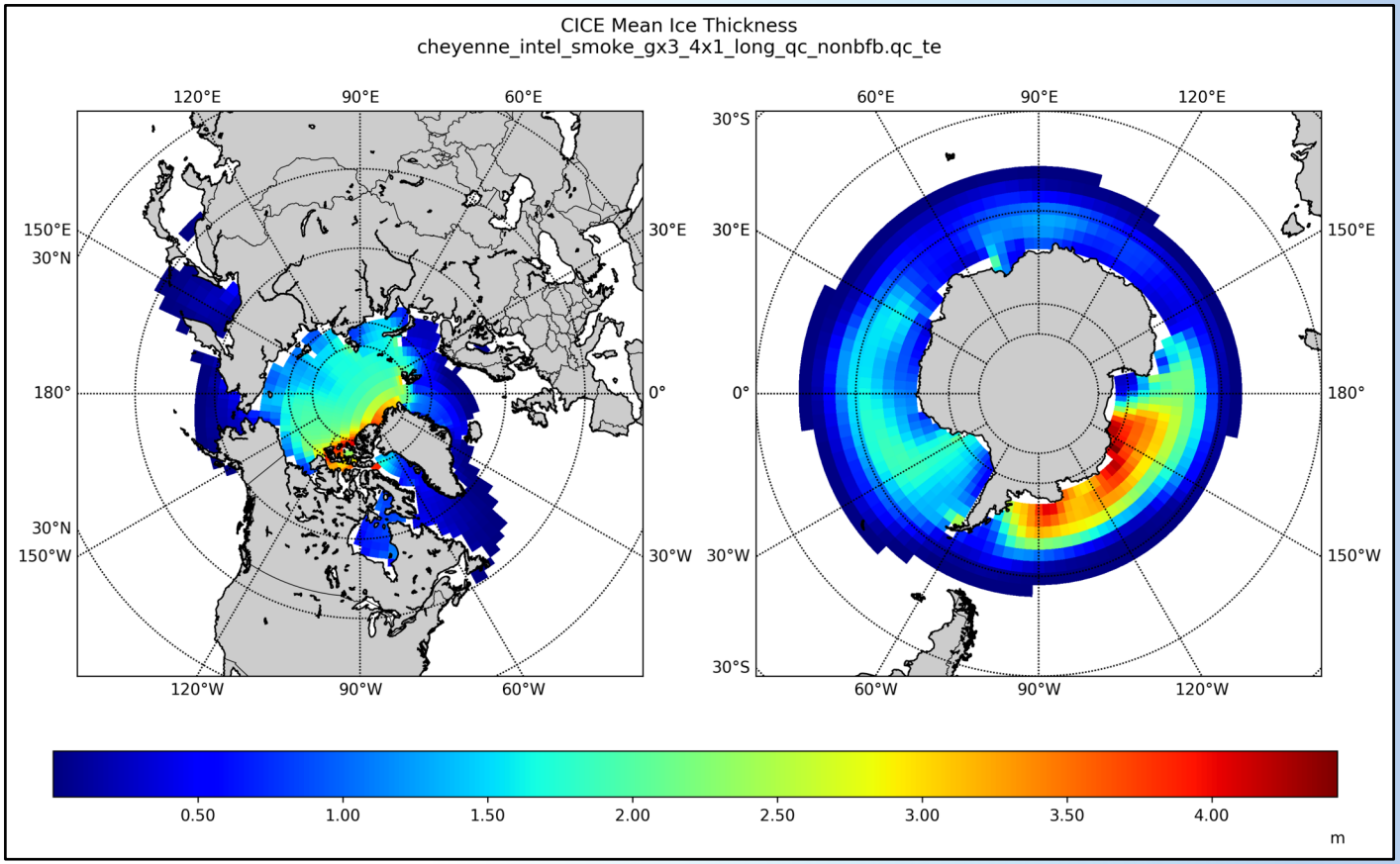
Creating map of the data:

Python script generates the following plots (.png)

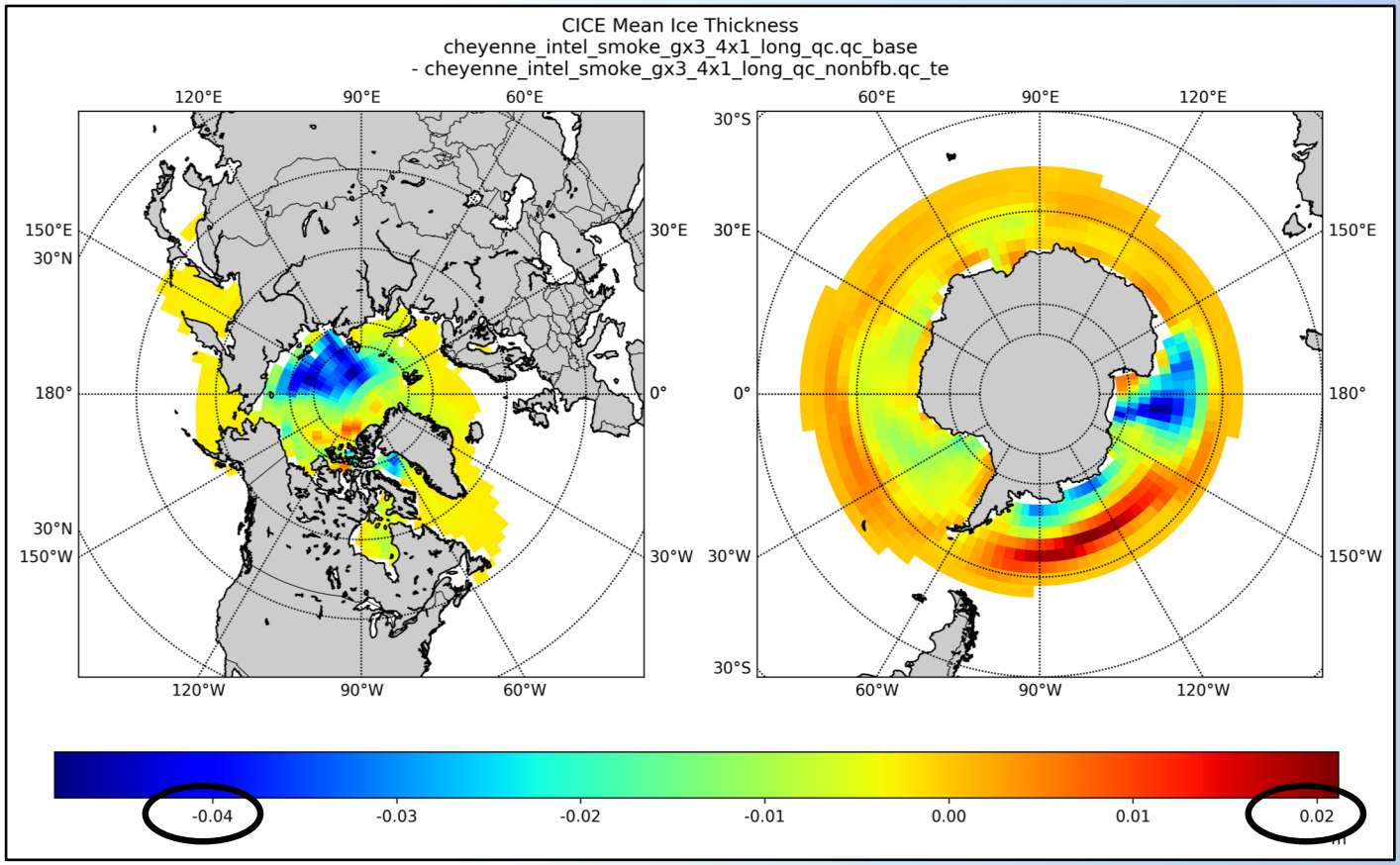
Map of Base test case results
Mean Ice Thickness (m)



Map of nonbfb test case results
Mean Ice Thickness (m)



Map of diff between base and nonbfb



Note small differences

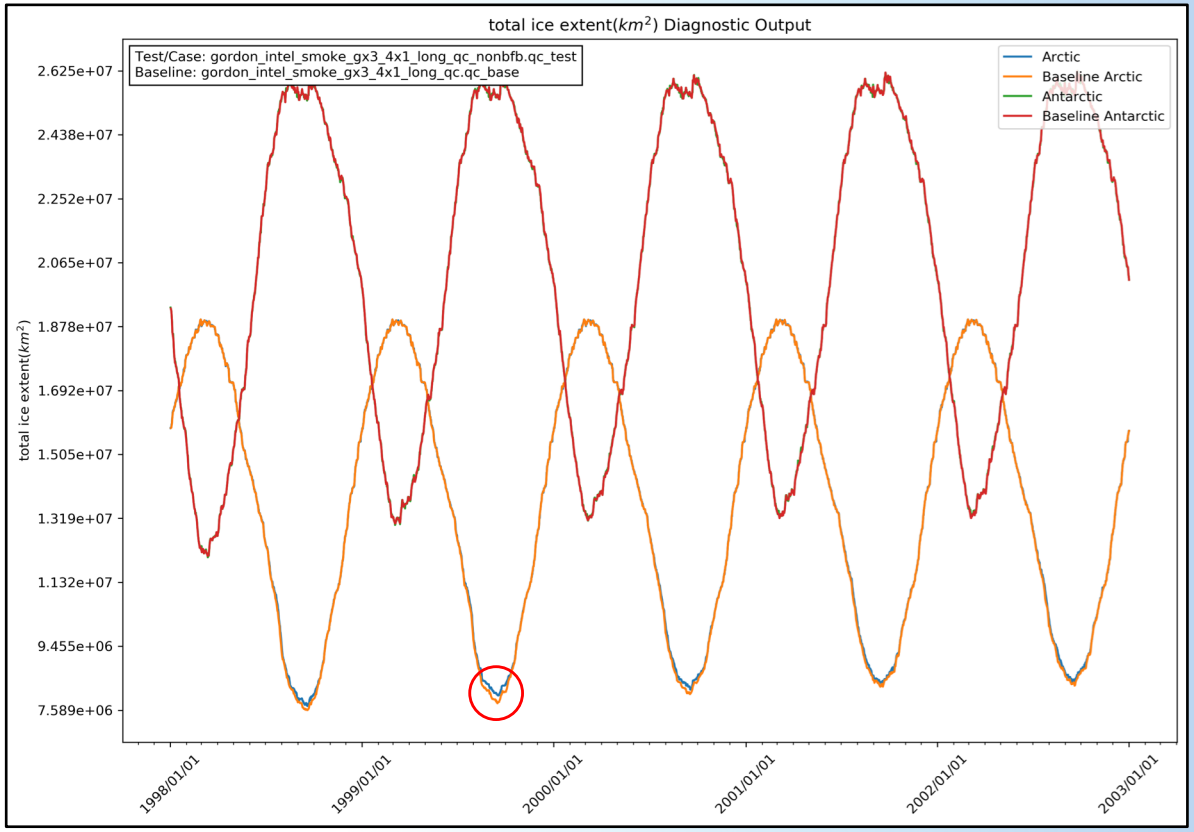
Quality Control Test Passed

Overlaying model results from 2 test cases on 1 plot

- Another tool available to examine model results is found in ./configuration/scripts/QC/**timeseries.py**
- This allows two different sets of model results to be plotted on the same figure

$ ./timeseries.py /path/to/test/log –bdir /path/to/base/log

Base case and Non BFB (not climate changing) cases



Only small changes in ice extent are evident.

Performing QC Analysis with CICE (cont.)

Now we will perform QC testing with Case 1 and Case 4

Running QC test on the following directories:

/glade/scratch/rallard//CICE_RUNS/gordon_intel_smoke_gx3_4x1_long_qc.qc_base/history

/glade/scratch/rallard/CICE_RUNS/gordon_intel_smoke_gx3_4x1_alt02_long_qc.qc_fail/history

Number of files: 1825

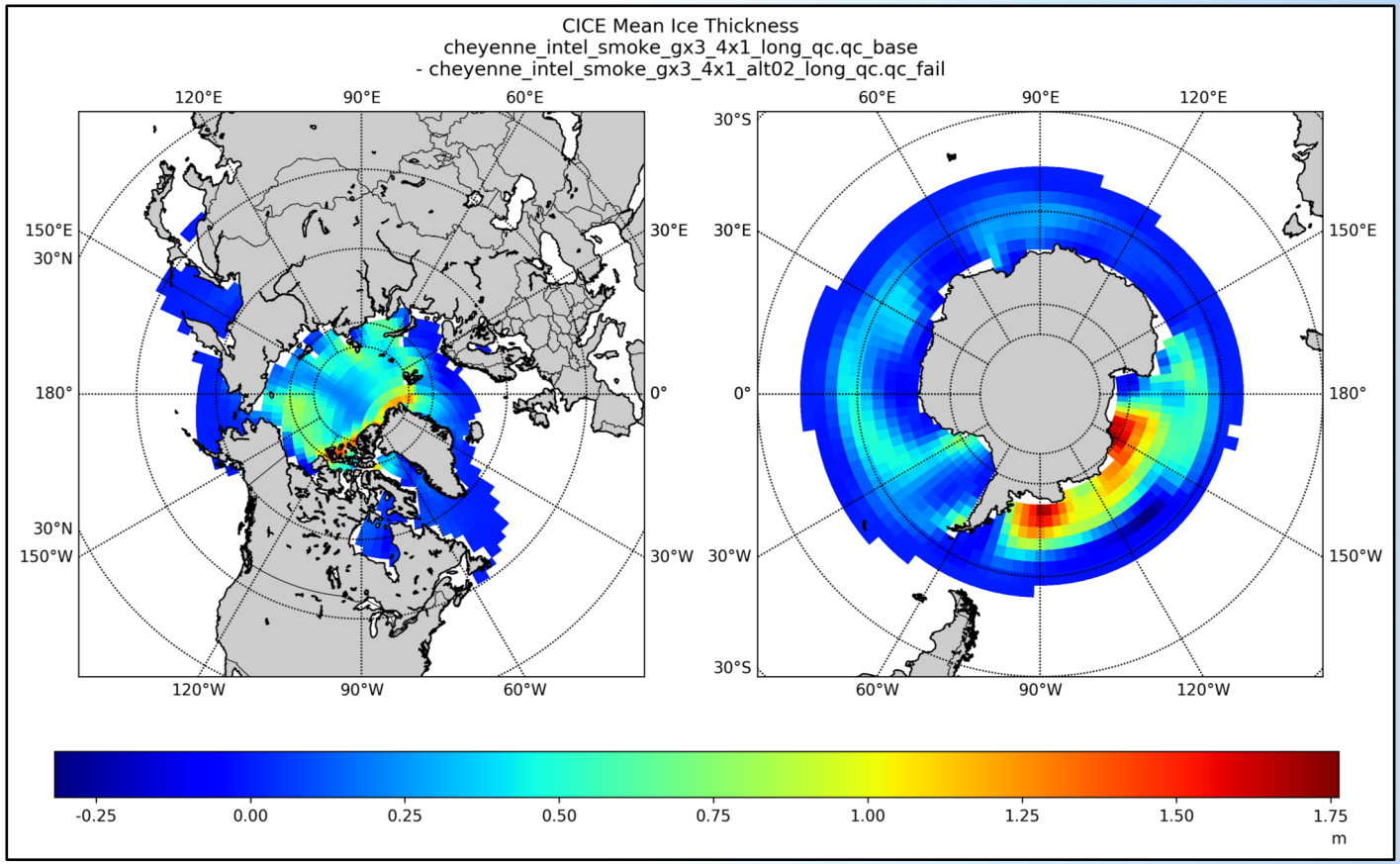
2 Stage Test Passed

Now check the spatial patterns of ice thickness from paired simulations to check if they are highly correlated and have similar variance.

Quadratic Skill Test Failed for Northern Hemisphere (S_{crit} < 0.99)

Quadratic Skill Test Failed for Southern Hemisphere (S_{crit} < 0.99)

Map of diff between base and nonbfb (climate changing)



Note significant differences found in northern and southern hemispheres

Quality Control Test Failed

Summary

- New CICE contributions (including Icepack) require QC testing to ensure that code modification (or additions) do not change the physics of existing model configurations when switched off.
- The CICE Consortium provides software tools to assess that code is either BFB (no QC test required), or requires additional testing (*2-stage paired thickness test, quadratic skill compliance test*) to determine if the new code does not produce significantly altered simulated ice volume using previous model configurations.
- If code does not pass the QC test(s), then further investigation will be required by the code developer.