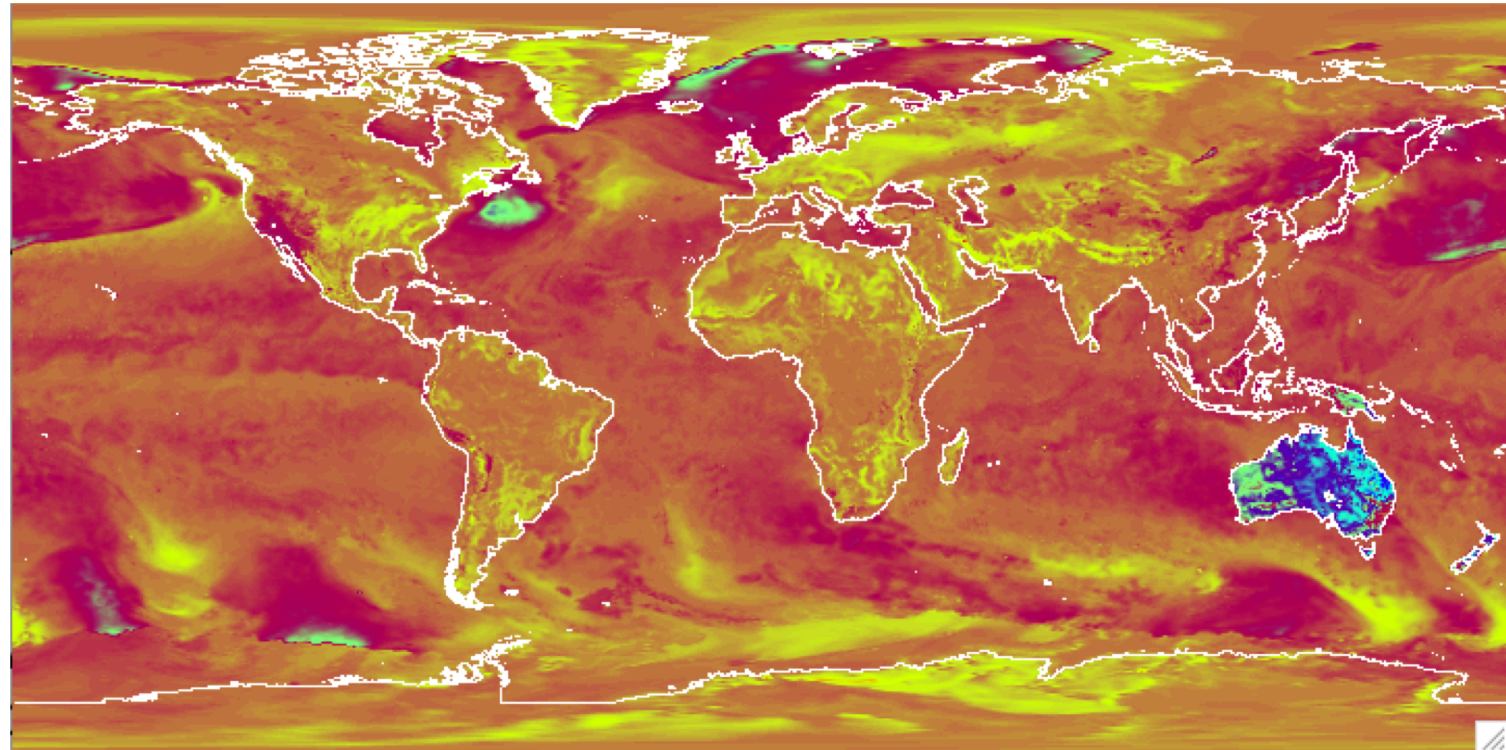


Quantifying “these results should be almost the same”;
For differences:
How BIG is TOO Big

Dave Gill (MMM/NCAR)

- Motivation
 - What is the problem we are trying to solve
 - How did we try to solve it
 - What looks promising
- Quick Review of ANOVA
 - Ratio of variabilities
 - Factors
- Results
 - Interpreting ANOVA table
 - Getting the p value
- Summary



Motivation: What is the problem we are trying to solve

- A modification is introduced into the built code, and the before vs after results are no longer bit-wise identical
- But we expect them to be pretty darn close (PDC)

New compiler
version, change
optimization

Different
compiler

Add LSB
perturbations
at IC

Hardware
differences
(CPU vs GPU)

Re-organized
loop, but
algebraically =

Single vs
double
precision

Motivation: What is the problem we are trying to solve

- A modification is introduced into the built code, and the before vs after results are no longer bit-wise identical
- But we expect them to be pretty darn close (PDC)
- We want to know if the fields are close enough (CE)
- How do we objectively state PDC is CE, and when PDC is not CE?

Motivation: What is the problem we are trying to solve

For those systems with bounded solutions, it is found that nonperiodic solutions are ordinarily unstable with respect to small modifications, so that slightly differing initial states can evolve into considerably different states.

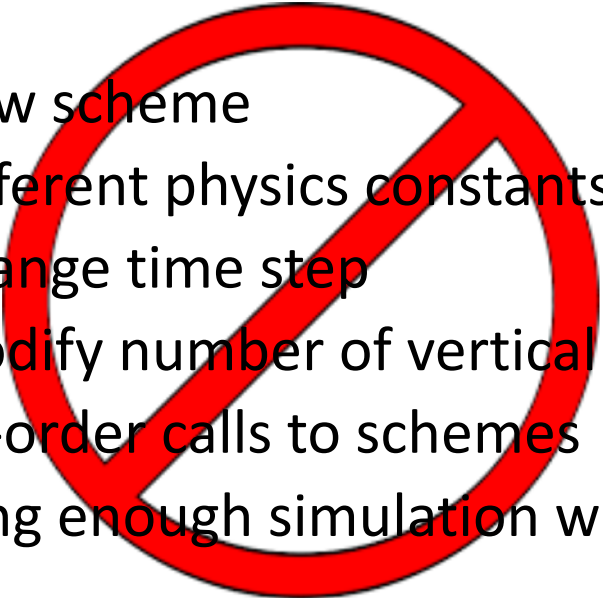
Lorenz, 1963

- Eventually, PDC will never be CE



Motivation: But *not* trying to solve this

- Can I say that the simulations are “similar” if I make substantive semantic changes? **WE ARE NOT DOING ANY OF THIS**

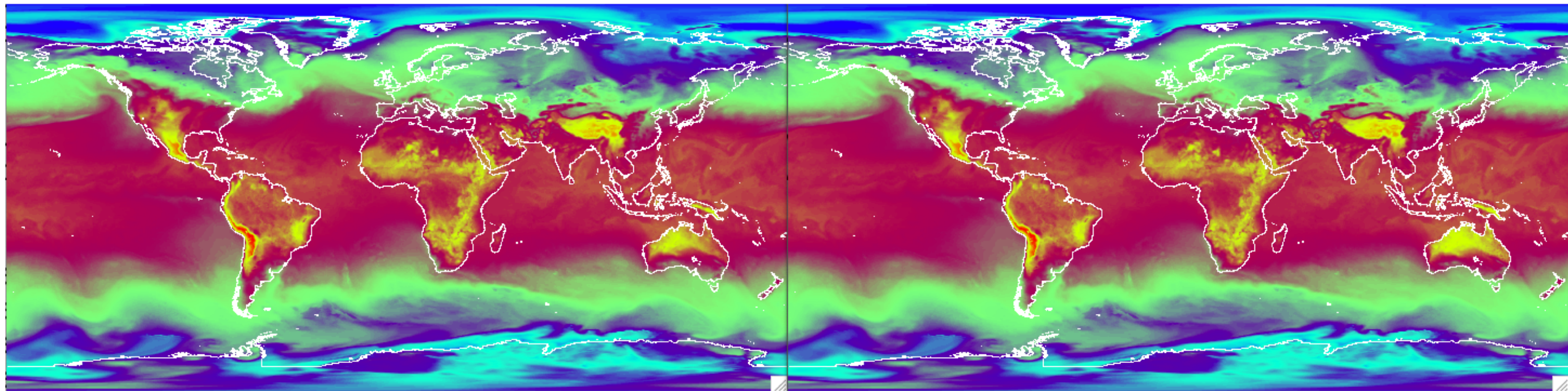
- 
- New scheme
 - Different physics constants
 - Change time step
 - Modify number of vertical levels
 - Re-order calls to schemes
 - Long enough simulation where path through other parts of code diverges

Motivation: How did we try to solve it

- Using a fully spun-up case from a restart file, look at the first time step with the two different set ups (either ICs or compiler related).
- We chose a wind field, thermal field, and a moisture field (any changes in one has to impact the others).
- Global domain gives us broad categories for free *(but we needed to identify categories; as a conglomerate is not helpful, even visually)*
 - Day vs night
 - Polar vs mid latitude vs equatorial regions
 - Marine vs land
 - Desert vs mountains

Motivation: How did we try to solve it

- An example of two low-level theta fields we want classified as different
- Often, one is able to visually detect unacceptable differences (not CE)
- This is unlikely to be the case when dealing with PDC



Motivation: How did we try to solve it

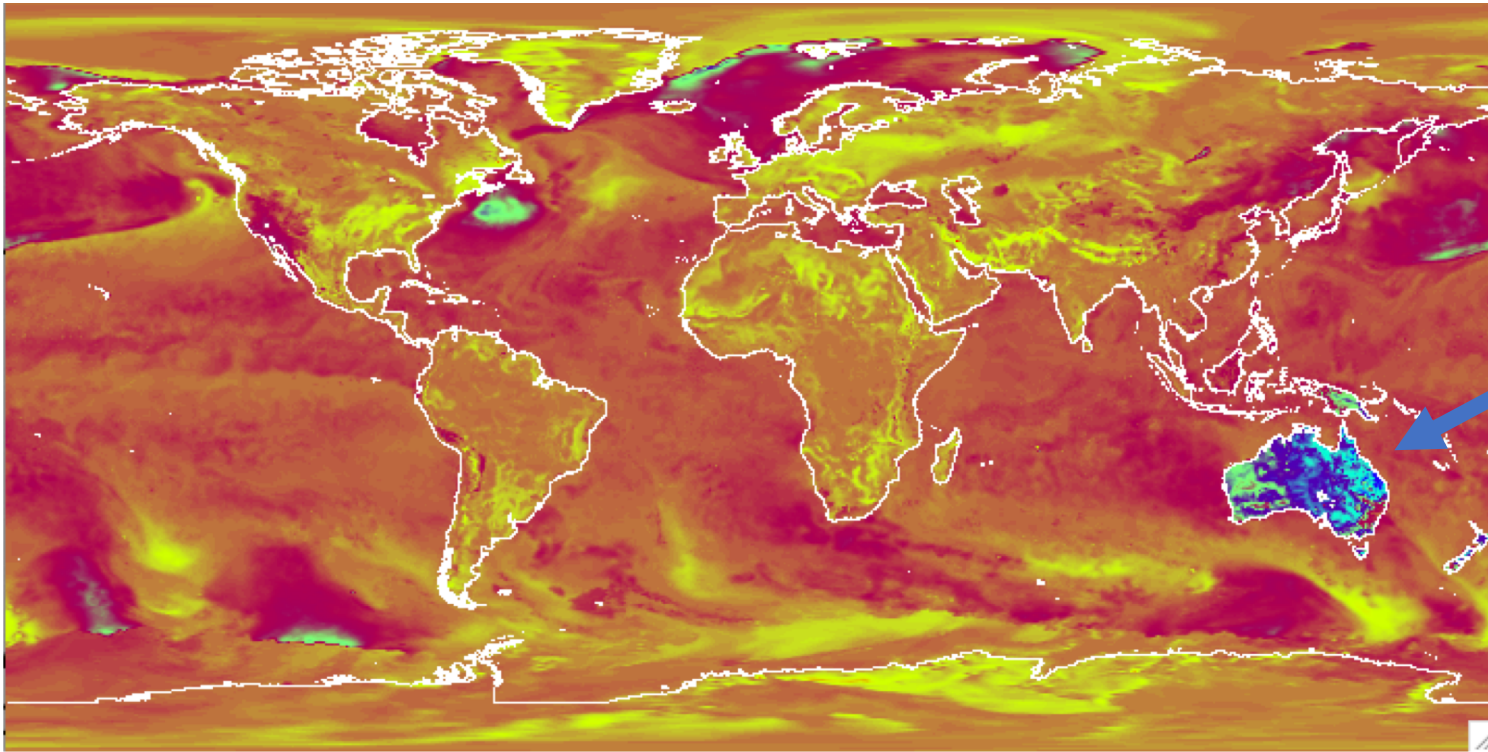
- Let's allow statistics to help us out of this quagmire
- Top values refer to left figure on previous slide, bottom values refer to right figure on previous slide
- Visually unable to detect differences; also, simple statistics => similar

	L2 norm	MAX	MIN	MEAN	STD
Qv =	+1.592540997467e+01	+2.455584891140e-02	+6.107344330154e-19	+2.678549848497e-03	+4.524076357484e-03
T =	+1.245417557043e+06	+9.468150024414e+02	+2.321143951416e+02	+3.875430603027e+02	+1.373361206055e+02
U =	+6.351148759083e+04	+1.062478713989e+02	-1.059075851440e+02	+1.342984195799e-02	+1.210562896729e+01
Qv =	+1.593991441577e+01	+2.428684942424e-02	+6.107337092323e-19	+2.680379431695e-03	+4.528562072664e-03
T =	+1.245418714720e+06	+9.468149414062e+02	+2.317701568604e+02	+3.875434875488e+02	+1.373358612061e+02
U =	+6.351069150938e+04	+1.062963562012e+02	-1.059652023315e+02	+1.341055892408e-02	+1.210547351837e+01

- Fiddlesticks

Motivation: How did we try to solve it

- Since we are comparing two widgets, how about the field difference?



In the biz, this is what we call a “clue”

Motivation: How did we try to solve it

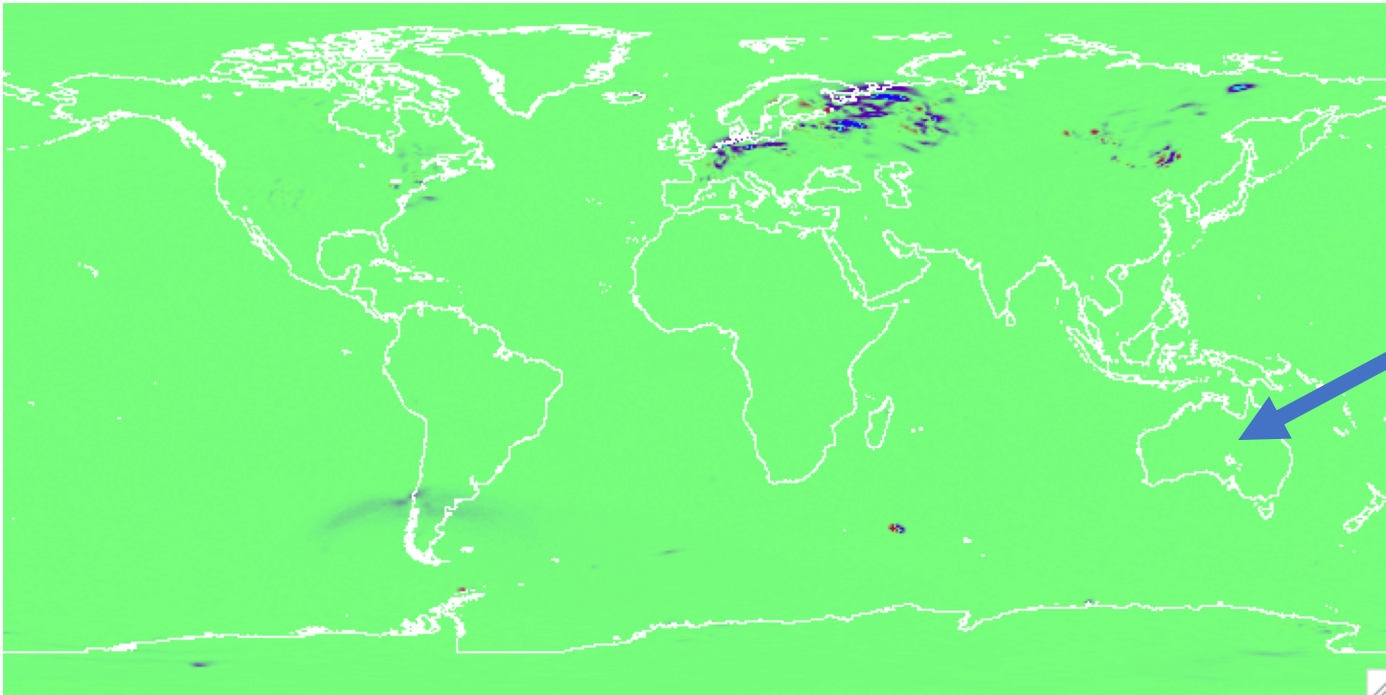
- Compute the same statistics with the difference fields

	L2 norm	MAX	MIN	MEAN	STD
Qv =	+1.328983029721e-01	+2.301060594618e-03	-4.858909174800e-03	-1.350821548840e-06	+3.485638808343e-05
T =	+1.449229523208e+02	+3.641113281250e+00	-3.607116699219e+00	-3.423761809245e-04	+3.803715854883e-02
U =	+2.497588993660e+02	+5.769649505615e+00	-6.485530853271e+00	+1.729066025291e-05	+6.555554270744e-02

- Globally, the difference field has mean = $O(10^{-4})$, STD = $O(10^{-2})$
- Wind and moisture are not much help either
- Again: Fiddlesticks

Motivation: What looks promising

- Australia (and NZ) were notable exceptions in the sfc diff



In the biz, this is also what we call a “*clue*”

- But not near the model top

Motivation: What looks promising

- With a number of other plots, we saw a clear pattern regarding differences for physics. They were related to:

- Land type
- Latitude
- Day night
- Altitude
- Simulation time



These we can
isolate via
specific global
locations

- By just looking at entire domain's results as a singular sample
 - We are masking important signatures
 - If there are significant differences, we can't identify an underlying cause

Motivation: What looks promising

- We can compare two small samples with Student's T-test, with a typical $\alpha = 0.05$
- However, if we try to add more simultaneous tests (say n comparisons), more attributes are compared.
- Our TYPE I error becomes $1 - 0.95^n$, allowing in false positives.
- So – problem: physically we want to isolate and conduct quite a few comparisons, but statistically it is a bit problematic.

Quick review of ANOVA

- ANalysis Of VAriance (ANOVA) is a technique that is able to compare multiple populations, and importantly for us, multiple subgroups of those populations.
- Looking at these multiple groups, do they likely come from the same population (this is always the null hypothesis). This inference comes from evaluating the various means.
 - Can we say that $\mu_1 = \mu_2 = \mu_3$
- This evaluation looks at the variance of several distributions, using the F-statistic.

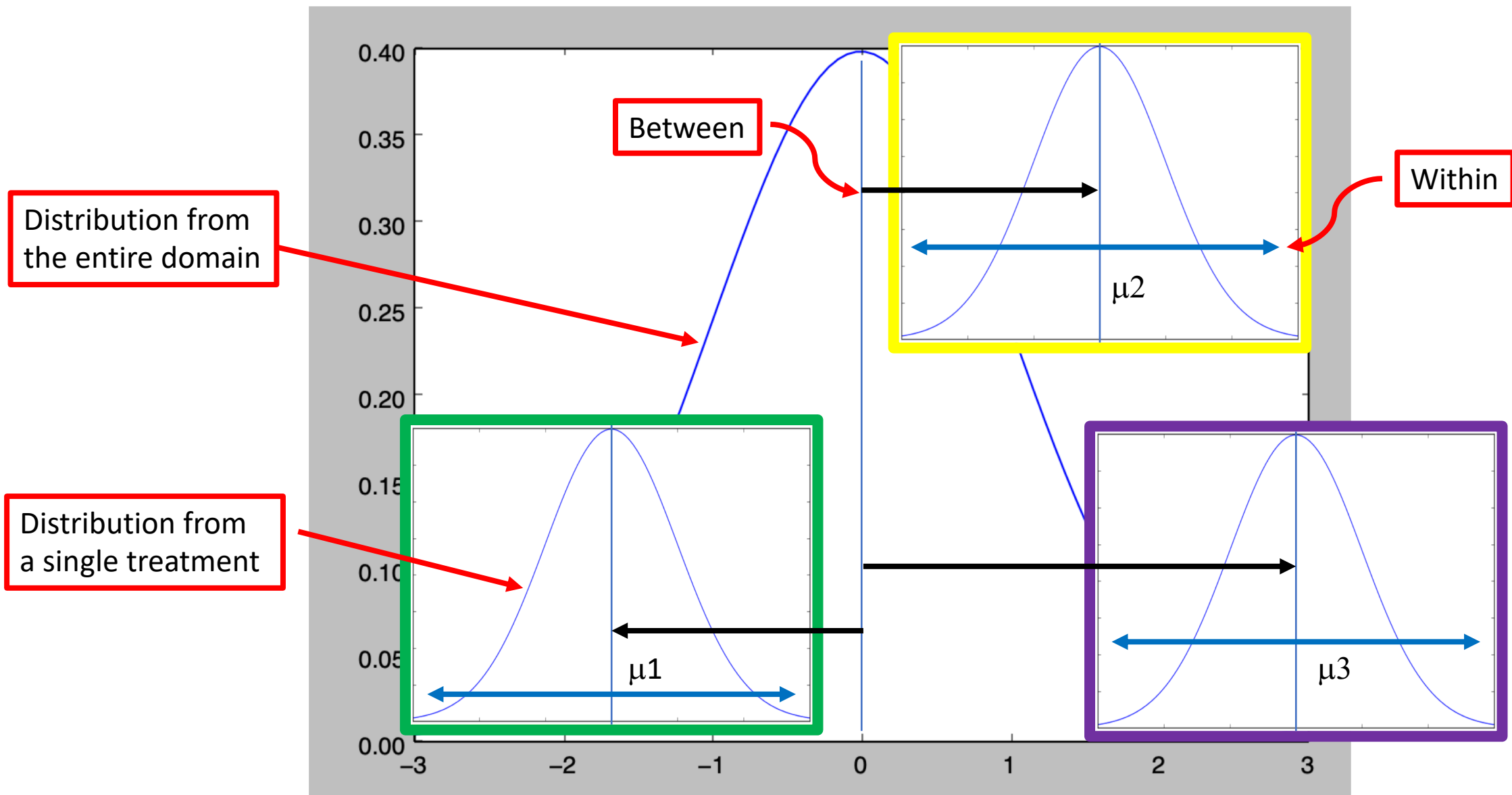
Quick review of ANOVA

- Our F-statistic is essentially a ratio of variabilities

BETWEEN the groups variability

WITHIN each group variability

- As this ratio approaches 0, the subgroup distributions are more *likely* identified with the overall distribution.
- As this ratio gets much larger than 1, the subgroup is increasingly *unlikely* to have been drawn from the same population.

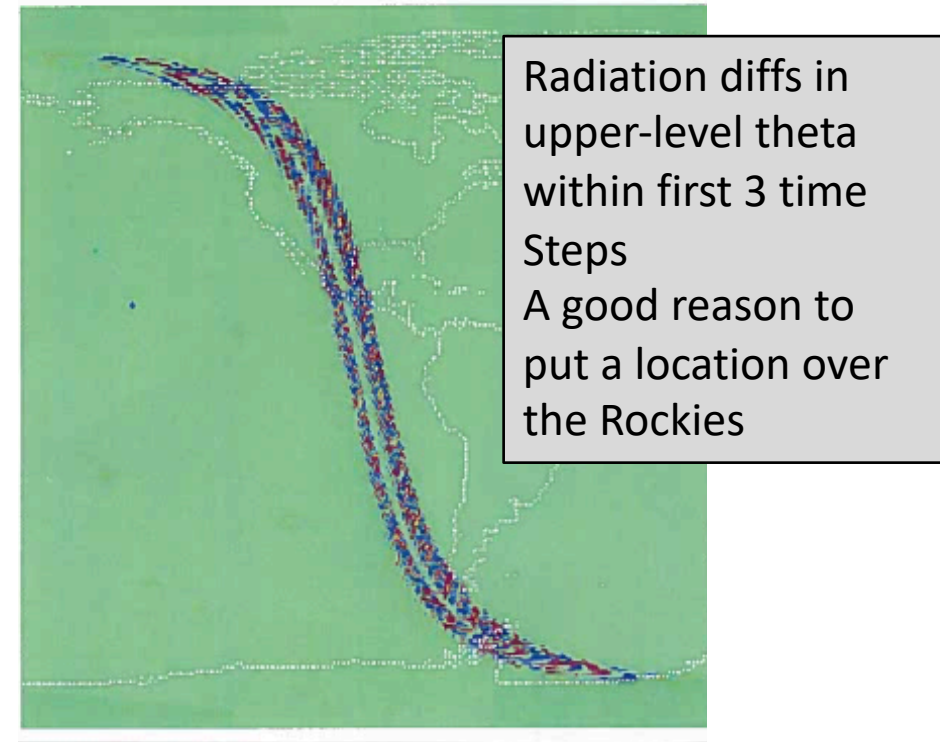
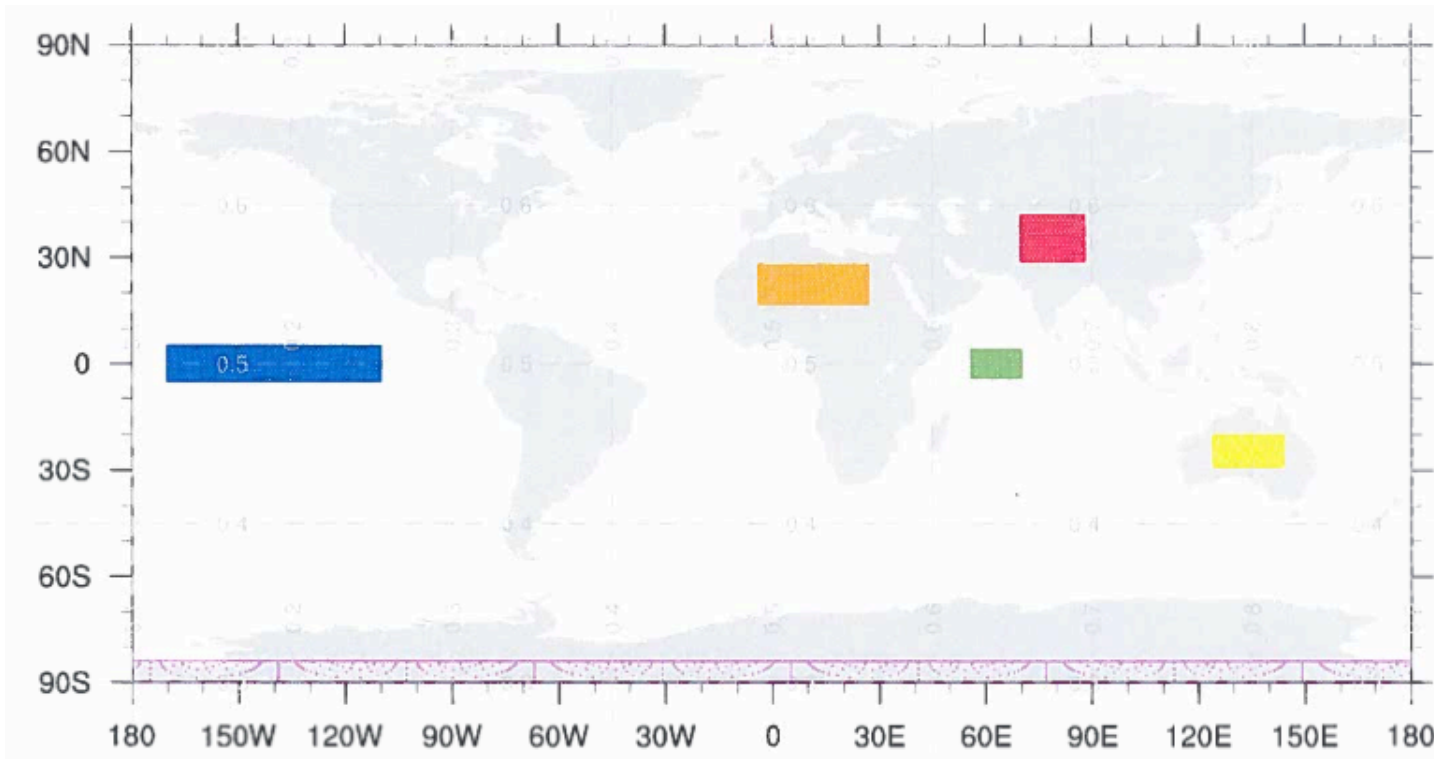


Quick review of ANOVA: Factors

- Consider mutually exclusive groupings of random model points
 - Locations
 - Time periods
 - Resultant “code” differences
- Each of the categories of groupings is a **FACTOR** (independent variable). Each factor has 2 or more **LEVELS**.

Quick review of ANOVA: Factors - Location

- Factor: Locations
 - Levels: Pacific, Indian, Himalaya, Sahara, Australia, Antarctica, Rockies



Quick review of ANOVA: Factors - Time

- Factor: Time of simulation beyond initialization (restart or cold start, both work well)
 - Levels: Time step = 1, 2, 3
- For an explicit example, with data from MPAS output validation files, where $dt = 6$ minutes

validation.2010-10-24_00.06.00_A.nc

validation.2010-10-24_00.12.00_A.nc

validation.2010-10-24_00.18.00_A.nc

Quick review of ANOVA: Factors – “Code Changes”

- Factor: Code changes may be from source modifications or differences in executable due to compile-time changes
 - Levels: Tests A, B, C, D
- A = Original
- B = No optimization
- C = Double precision
- D = Different PBL scheme

Quick review of ANOVA: Factors

- Each combination of factor levels is a **TREATMENT**
- All treatments have 20 randomly chosen data points from that specific location (geographical box)
- The multiple factors:
 - Remove **CONFOUNDING**, which is not allowing elimination of plausible explanations
 - Reduce variability within each treatment

Quick review of ANOVA: Factors

- Total treatments: 7 locations * 3 Time periods * #Code changes
 - We only care about a single H_0 : are the code changes a significant source of error
 - There are actually seven available H_0 !
- A=original, B=no optimization, C=double precision, D=different PBL
- First example is **A vs B vs C** and should be **A-OK**
- The second example is **A vs D**, should **fail**

RESULTS: A vs B vs C (orig vs no opt vs dbl prec)

Source		df	SS	MS	F Statistic
=====					
	Mean	0001	0.26650	0.26650	26.023
	LOCATIONS	0006	61.69340	10.28223	1004.022
	COMPILERS	0002	0.00012	0.00006	0.006
	TIMES	0002	0.03195	0.01597	1.560
	LOCATIONS x COMPILERS	0012	0.00124	0.00010	0.010
	LOCATIONS x TIMES	0012	8.50036	0.70836	69.169
	COMPILERS x TIMES	0004	0.00000	0.00000	0.000
	LOCATIONS x COMPILERS x TIMES	0024	0.00005	0.00000	0.000
	Error	1197	12.25854	0.01024	

Interpreting these values:
 $F_{stat} = \text{COMPILERS MS} / \text{Error MS}$

RESULTS: A vs B vs C (orig vs no opt vs dbl precision)

compiler comparison for theta

Input, F-statistic: 0.006045622805382375

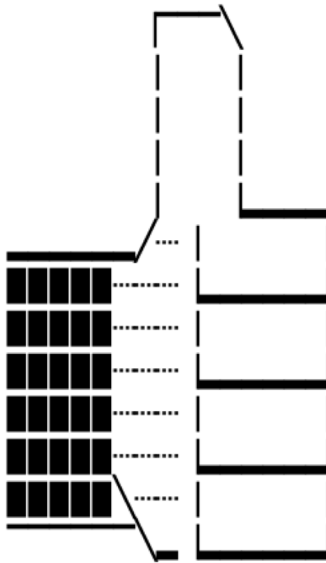
Input, df factor: 2

Input, df error: 1197

p-value probability = 1.0 means 100% reject null hypothesis that means are same

p-value probability = 0.006027354449653595

We are pretty darn confident that our comparisons are OK



RESULTS: A vs D (orig vs diff PBL)

Source		df	SS	MS	F Statistic
=====					
<div> <p>We do not care about LOCATIONS or TIMES. We know they have an impact. Explicitly including those factors removes confounding.</p> <p>Also, no interest in any interaction</p> </div>	Mean	0001	15.89986	15.89986	848.482
	LOCATIONS	0006	185.59000	30.93167	1650.642
	COMPILERS	0001	12.76186	12.76186	681.026
	TIMES	0002	0.04354	0.02177	1.162
	LOCATIONS x COMPILERS	0006	54.55443	9.09240	485.208
	LOCATIONS x TIMES	0012	0.94872	0.07906	4.219
	COMPILERS x TIMES	0002	0.06236	0.03118	1.664
	LOCATIONS x COMPILERS x TIMES	0012	5.11211	0.42601	22.734
	Error	0798	14.95386	0.01874	

RESULTS: A vs D (orig vs diff PBL)

compiler comparison for theta

Input, F-statistic: 681.0256031623537

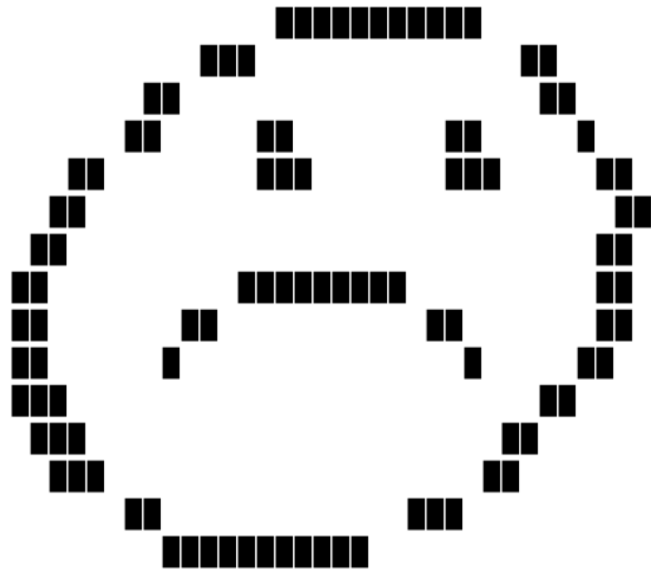
Input, df factor: 1

Input, df error: 798

p-value probability = 1.0 means 100% reject null hypothesis that means are same

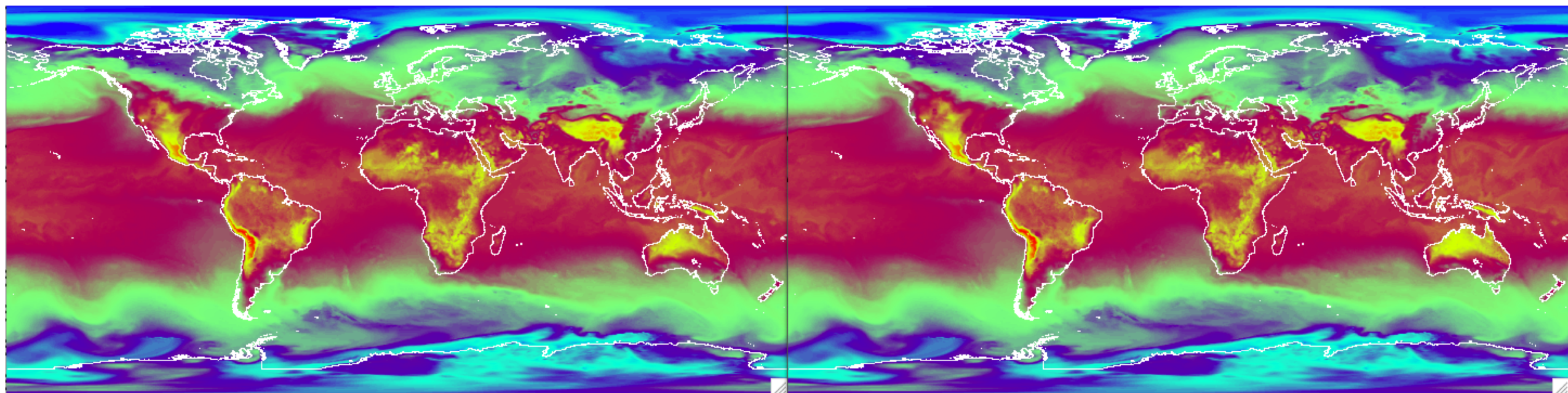
p-value probability = 1.0

There are DEFINITELY problems with the GPU comparison



Summary

- We defined exemplars constituting “similar” and “different”
- With ANOVA, good experimental design, and significance *tuning* we are able to effectively identify “similar” and “different”

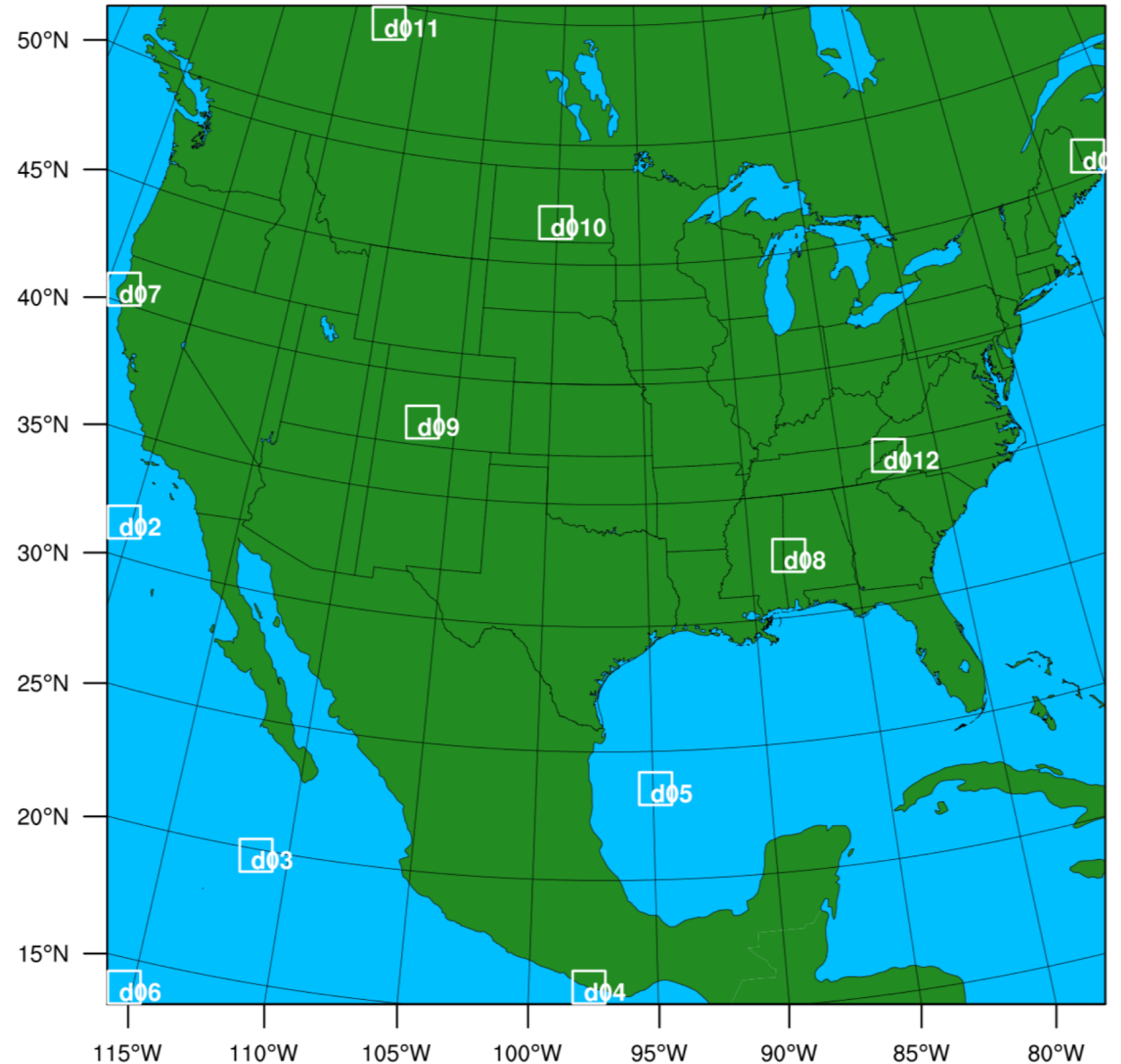


Summary: FACTORS

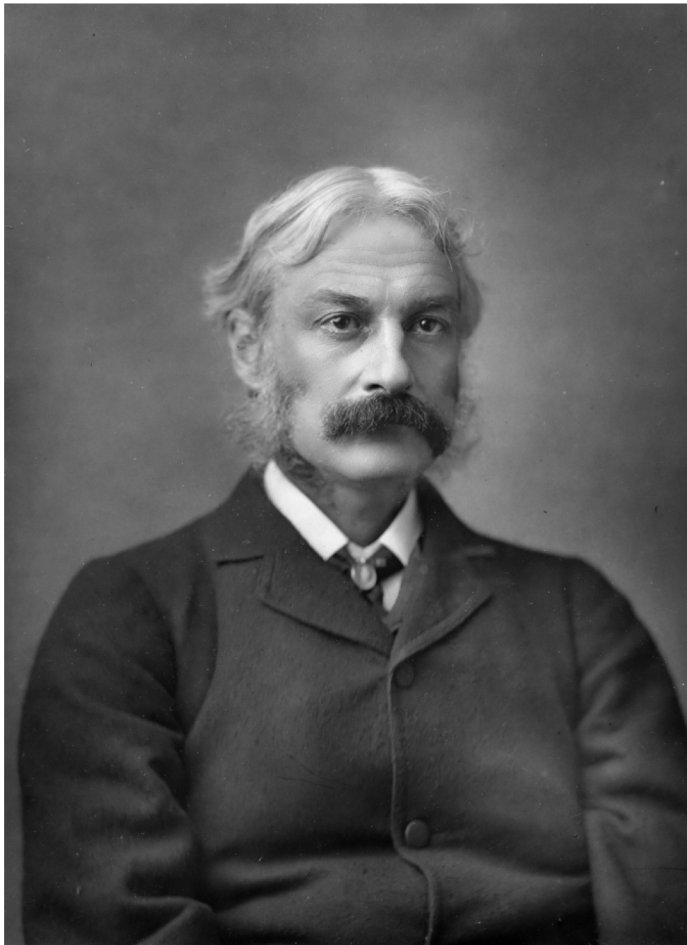
- Enough FACTORS need to be included to remove **CONFOUNDING**
- Statistics from extra factors and all interactions are ignored
 - Additional factors are simply for removing confounded explanations
- For global models, specific **lat/lon locations** are easily defined
- More **variables** may need to be included: skin temperature, jets, soil conditions, various fluxes, etc
 - For example, upper-level U is only mildly impacted in one time step from PBL
- Only **a few time steps** of model simulation are required
 - Initialization of the model from cold start ICs or restart both perform OK

Summary: Regional, too!

- It is just more tedious to choose locations for a regional domain
- Want a mix of
 - Day vs night
 - Water vs land
 - North vs south
 - Mountain vs flat
 - Interior vs boundary
 - Inflow vs Outflow



Summary



I shall try not to use statistics as a drunken man uses lamp-posts, for support rather than for illumination

Andrew Lang

Scottish novelist and folklorist

1844 - 1912

Extra slides

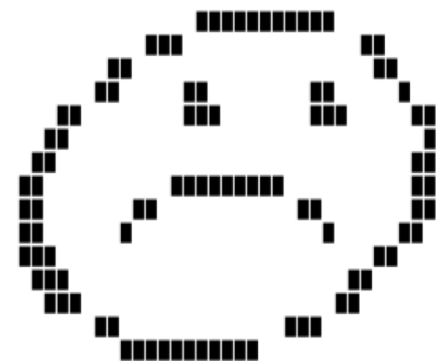
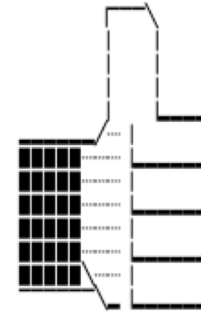
https://github.com/davegill/TWC_VALIDATION

Summary: FACTORS

- Factors are the independent variables
- Different from correlations, the independent variables for ANOVA are categorical: time step #3, Antarctica, Intel compiler 19.0.1, etc
- Dependent variables are the measurements
- Measurements are quantitative values
- All of the computations are run with the dependent variables, where the collections of dependent variables are chosen from the different independent variable treatments

Summary: PDC, but CE??

- Per field symbol assignment:
 - Thumbs up: “field similar”, $p \leq 0.1$
 - Frown: “field different”, $p \geq 0.9$
 - Question mark: “Hmmm”, $0.1 < p < 0.9$
- Simulation symbol interpretation:
 - ALL fields show thumbs up: “similar”
 - ANY field shows a frown: “different”
 - ELSE: “We need to look more closely”



Summary: Randomization and Sample Size

- Currently use 20 random points from each location (geo-box)
- Used several random choices of 20 points, all gave OK results
- Used 10 – 30 random points, all gave OK results
- Tried 5 random points per geo-box, results sensitive to point selection choices
- Stayed away from larger collections of points since we could get statistical significance but not practical significance (much of our area of coverage could have low variability)

Summary: How I do this for MY model

- Takes approximately 1 minute to generate actionable output with ANOVA script
 - My time goes up linearly with # global domain points due to my searching algorithm for lat/lon geo-box
- 3-way ANOVA in R is available online for download
 - Plenty of online examples with which to vet the code and data ingest
- Python script to compute p-values from F-statistic and dfs is online (it is where I got mine)

Summary: How I do this for MY model

- Experimental design
 - Causal relationships between independent and dependent variables
 - Control impact from factors outside of the identified independent variables (remove confounding)
 - Reduce variability within each treatment, make detection of differences easier (single levels of fields, differing locations, individual fields, time levels, etc.)

Summary: How I do this for MY model

1. Read exemplar simulation, test simulation, and IC
2. Find cell locations inside of requested geo-boxes
3. Diff simulation data (exemplar – IC, test – IC) within those boxes, for requested fields, for requested times
4. Randomly choose requested # of points within each geo-box
5. Run 3-way, balanced ANOVA
6. Determine probability of rejecting H_0