Machine Learning for Model Calibration

M2LInES Team Meeting

V. Balaji

CIMES, Princeton University and NOAA/GFDL

8 August 2021

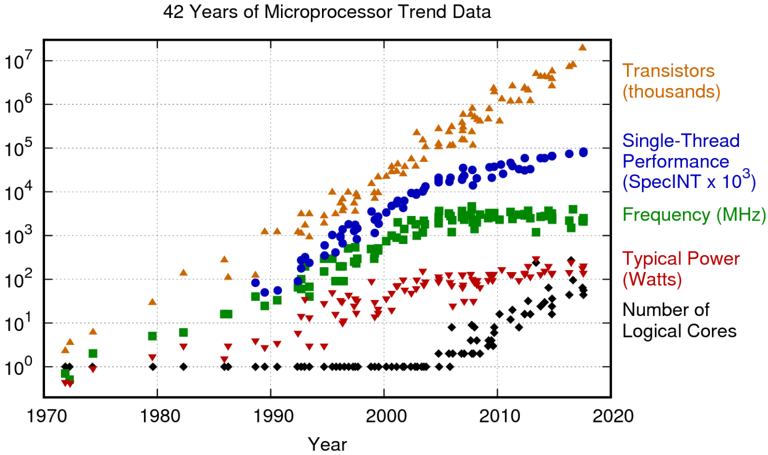
Outline

1. Computing technology is moving towards deep learning
2. Deep Learning for climate: no black boxes, no free lunches
3. Machine Learning for Model Calibration
    - Model calibration or “tuning”, a brief survey
    - HighTune: History matching
    - CliMA: Calibrate-Emulate-Sample
    - Calibrating coupled models with long timescales
4. Summary: prospects and challenges for model calibration using ML

Outline

1. Computing technology is moving towards deep learning
2. Deep Learning for climate: no black boxes, no free lunches
3. Machine Learning for Model Calibration
   - Model calibration or “tuning”, a brief survey
   - HighTune: History matching
   - CliMA: Calibrate-Emulate-Sample
   - Calibrating coupled models with long timescales
4. Summary: prospects and challenges for model calibration using ML

End of Dennard scaling

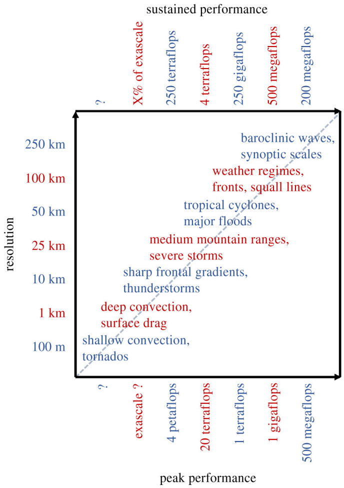


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

From 42 Years of Microprocessor Trend Data, courtesy Karl Rupp.

What can we expect at an exaflop?

Will exascale be the rescue? Neumann et al (2019).



Hypothesis: vastly reduced uncertainty at $\sim$1 km (see "digital twins", DestinE, NextGEMS, ...)

- ICON projects that a 1 km global model will run at 0.06 SYPD on "pre-exascale" technology: 17X improvement needed for 1 SYPD.
- This will be on 200,000 nodes (roughly 2xGaea).
- DECK: 1000 SY.
- A full suite of hindcasts for seasonal forecasting: 10,000 SY.
- Ocean state needed for seasonal and beyond prediction as well!

All algorithms are not created equal

- Real codes often gated by memory bandwidth.
- Roofline model:

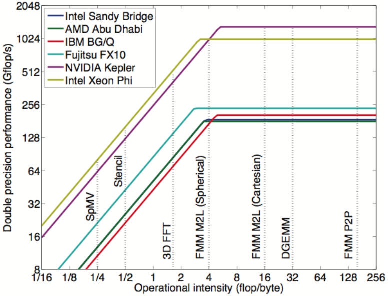
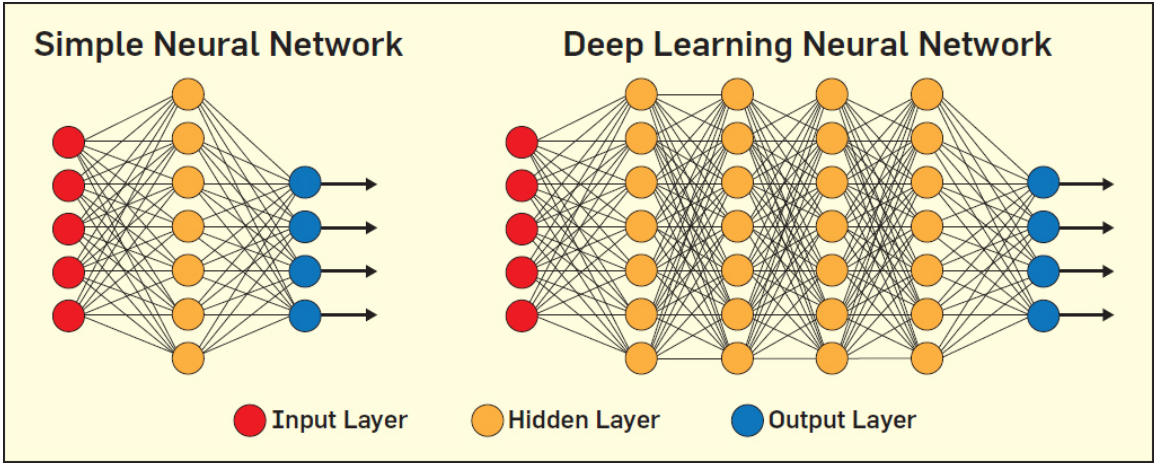


Figure courtesy Barba and Yokota *SIAM News* 2013.

Deep Learning

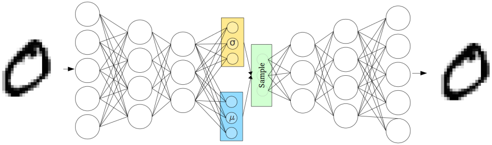


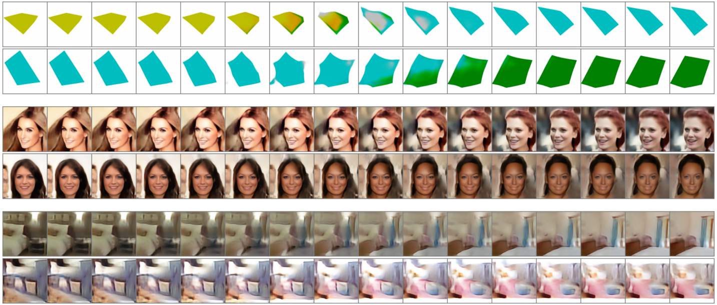
From Edwards (2018), ACM. Dense linear algebra with high operation intensity, data-intensive.

Outline

1. Computing technology is moving towards deep learning
2. Deep Learning for climate: no black boxes, no free lunches
3. Machine Learning for Model Calibration
   - Model calibration or “tuning”, a brief survey
   - HighTune: History matching
   - CliMA: Calibrate-Emulate-Sample
   - Calibrating coupled models with long timescales
4. Summary: prospects and challenges for model calibration using ML

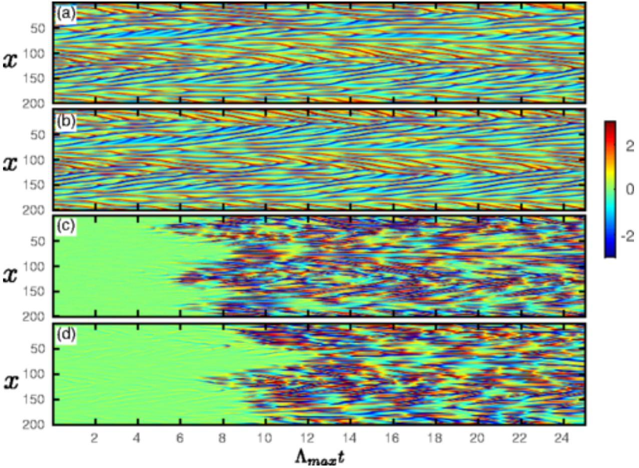
The ML approach: finding the essence





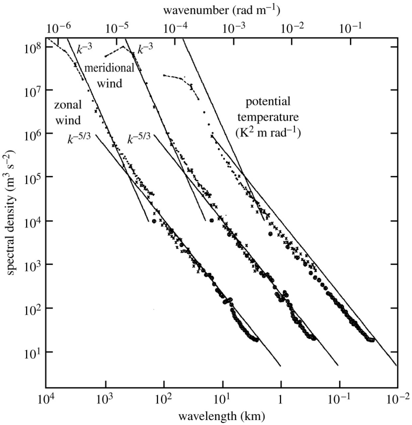
From “features” make new instances that capture the essence Angles and Mallat (2018)

Model-free prediction or model enhancement?



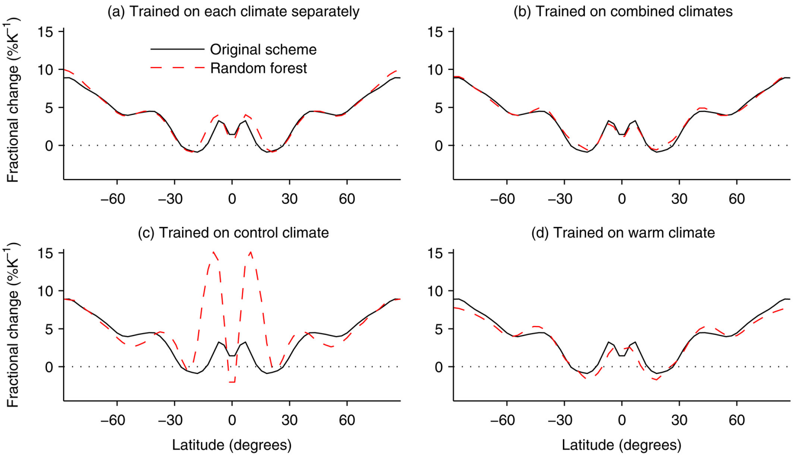
From Pathak et al, PRL (2018), *Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach*. See also Patel et al (2021).

No separation of "large" and "small" scales



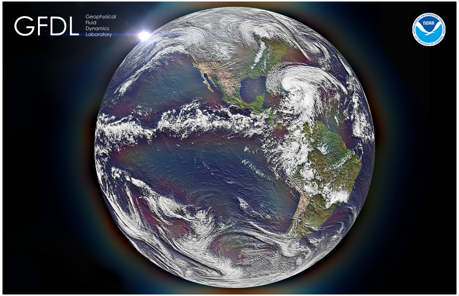
Nastrom and Gage (1985). More model fidelity, more complexity over time in small scales ("physics"). The backscatter idea (Jansen and Held 2014) provides an energetically consistent framework for SGS.

Replacing a parameterization with DL

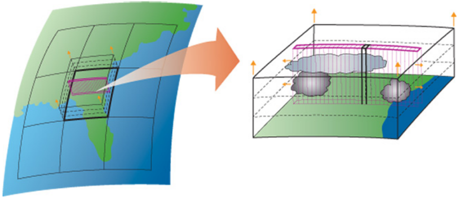


From O'Gorman and Dwyer (2018). Limitations of training on short non-stationary time series. See also Dixon et al (2016).

Scale separation: superparameterization



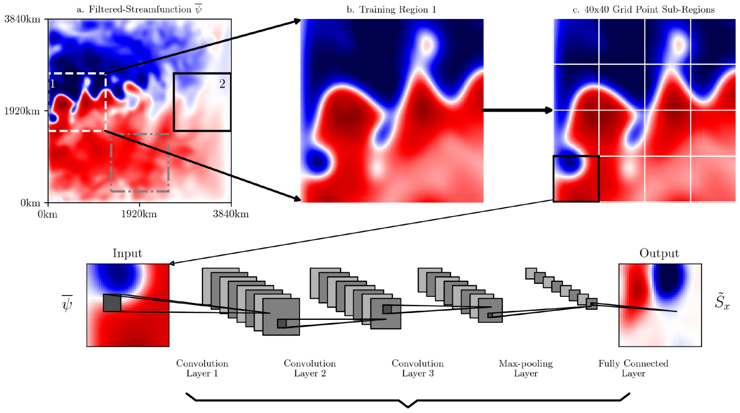
(Courtesy: S-J Lin, NOAA/GFDL).



(Courtesy: D. Randall, CSU; CMMAP).

- Global-scale CRMs (e.g 7 km simulation on the left) and even super-parameterization using embedded cloud models (right) remain prohibitively expensive.
- Can we learn the statistical aggregate of small scales? See Schneider et al 2017, Gentine et al (2018), O'Gorman and Dwyer (2018), Bolton and Zanna (2019), ...

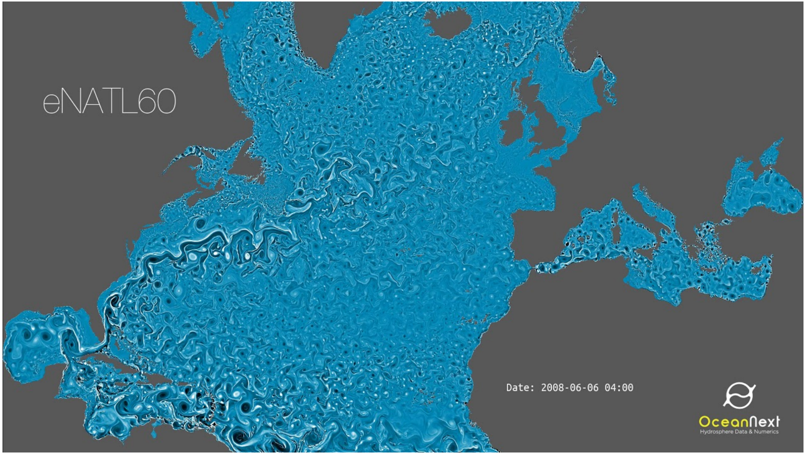
Learning sub-gridscale turbulence



Neural network $\tilde{S}_x = f_x(\overline{\psi}, \mathbf{w}_1)$, trained to minimize loss $L \propto (S_x - \tilde{S}_x)^2$.

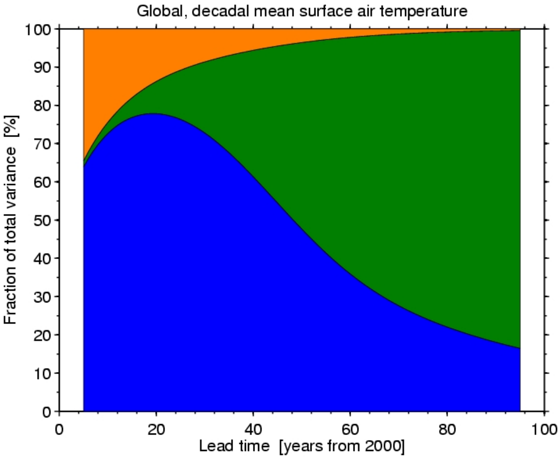
Fig 1 from Bolton and Zanna (2019).

Coarse-graining without scale separation



eNATL60 dataset courtesy Julien le Sommer and collaborators. Can we assume a structure for learning. e.g “GM+E” Bachman 2019. See Sommer et al AGU 2019.

Science requires going beyond observations

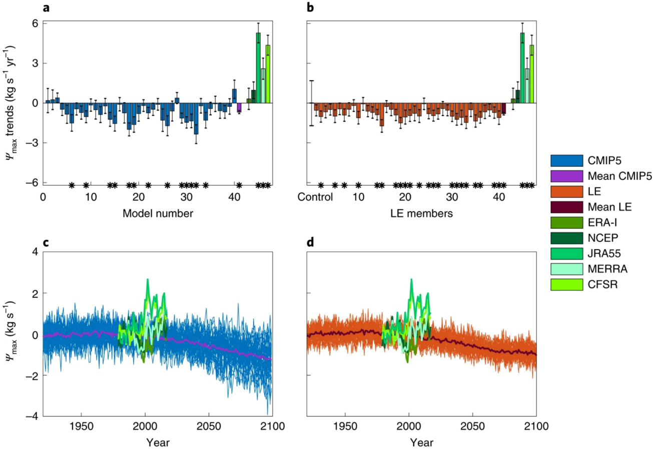


Sources of uncertainty in weather and climate simulation:

- *chaotic uncertainty* or internal variability
- *scenario uncertainty* dependent on policy and human actions.
- *structural/epistemic uncertainty* or imperfect understanding.

Models must also generate counterfactual values! From Hawkins and Sutton (2009).

Models or observations?



Hadley cell strength is likely correct in models and not in “observations”!
From Chemke and Polvani (2019).

Outline

1. Computing technology is moving towards deep learning

2. Deep Learning for climate: no black boxes, no free lunches

3. Machine Learning for Model Calibration
   - Model calibration or “tuning”, a brief survey
   - HighTune: History matching
   - CliMA: Calibrate-Emulate-Sample
   - Calibrating coupled models with long timescales

4. Summary: prospects and challenges for model calibration using ML

ML for model calibration: the sales pitch!

- Models, even "seamless" ones, may be configured or calibrated differently for different problems (e.g forecast horizons).
- Each problem carries an implicit cost function by which a model configuration is declared suitable.
- Models do not converge cleanly with resolution: much unresolved physics is not yet "scale-aware".
- Computation alone is not going to make the problem go away (not everyone agrees...)
- Important new constraints on models from observations (new generation of satellites, Argo...)
- While data science is a misnomer (what is non-data science?) the convergence of computation and statistics that we call ML provides paths forward toward seamlessness: traceable hierarchies of scale, *Charney's ladder*

Model calibration

Model calibration or “tuning” consists of reducing overall model error (relative to some goal of modeling) by modifying parameters. In principle, minimizing some cost function:

$$C(p_1, p_2, ...) = \sum_1^N \omega_i \|\phi_i - \phi_i^{obs}\|$$

- Usually the p must be chosen within some observed or theoretical range $p_{min} \leq p \leq p_{max}$.
- “Fudge factors” (applying known wrong values) generally frowned upon (see Shackley et al 1999 on “flux adjustments”.)
- The choice of ω_i is part of the lab's “culture”. Cost also plays a role.
- The choice of ϕ_i^{obs} is also troublesome:
    - overlap between “tuning” metrics and “evaluation” metrics.
    - “Over-tuning”: remember “reality” is but one ensemble member...

See for example, Hourdin et al (BAMS 2017)

Sources of diversity among models

- Goals of modeling: understanding, prediction, mission.
- Scientific interests of model developers.
- History and genealogy of models.
- Approach to model construction and tuning.
- Computing and data constraints.

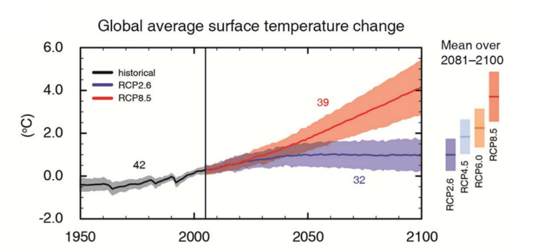
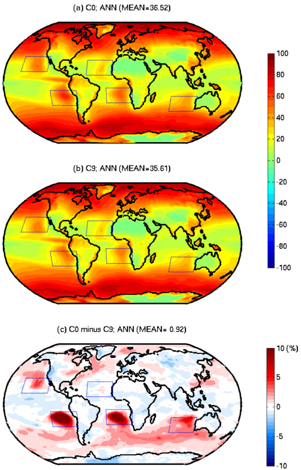
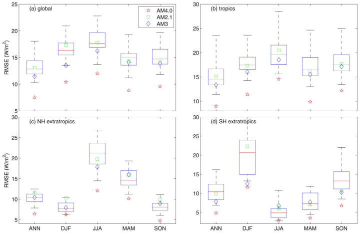
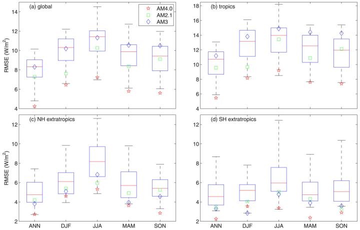


Figure SPM.7 from the IPCC AR5 Report. Global mean temperatures reduced by imposing CO_2 controls.

Example: the tuning of GFDL's AM4/OM4/CM4 models







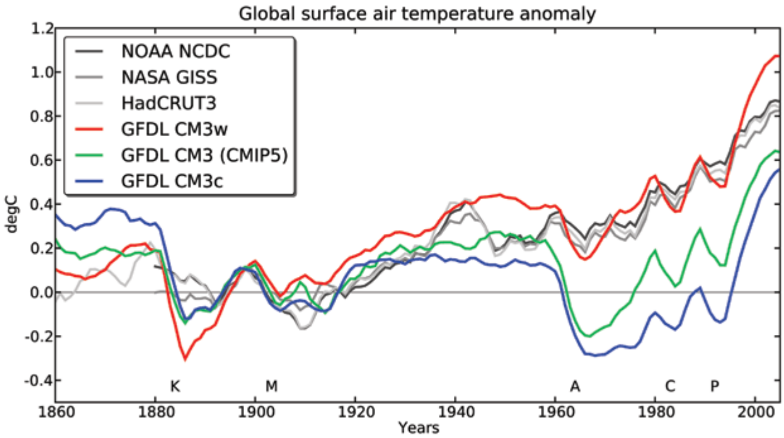
From Zhao et al (2018).

Example: the tuning of GFDL’s AM4/OM4/CM4 models

- The GFDL Global Atmosphere and Land Model AM4.0/LM4.0: 2. Model Description, Sensitivity Studies, and Tuning Strategies
- The GFDL Global Ocean and Sea Ice Model OM4.0: Model Description and Simulation Features: “We hypothesize that the development of a climate model is optimized only with close coordination across component model development.”
- Structure and Performance of GFDL’s CM4.0 Climate Model: “CM4.0 is sensitive to a number of features [...] much less apparent in uncoupled atmosphere/land simulations”
- Climate Sensitivity of GFDL’s CM4.0
- The GFDL Earth System Model Version 4.1 (GFDL-ESM 4.1): Overall Coupled Model Description and Simulation Characteristics

The JAMES special issue on GFDL’s “4 series” models. 50,000 SY of coupled models run during model development, 10,000 SY of “CMIP6 runs”.

Should we tune to get the 20th century?



Tuning reduces model bias without violating process fidelity (but poses a problem for validation). From Golaz et al 2013.

Parameter optimization, elimination, uncertainty quantification

Goal: explore parameter space of model while minimizing the use of expensive forward models.

- Parametric uncertainty vs structural uncertainty.
- A two stage process: process fidelity followed by global constraints.
- The choice of cost function.
- Metric weights and normalization.
- Do observations sample the space sufficiently?
- If models "higher on the ladder" are used for calibration, are they representative of all possible states? What the associated uncertainties?
- Internal feedbacks on multiple timescales, and compensating errors.

HighTune: Formulating the problem

$$\frac{\partial \mathbf{x}}{\partial t} = D(\mathbf{x}) + \sum_n (\mathcal{P}_n(\mathbf{x}, \lambda_n)$$

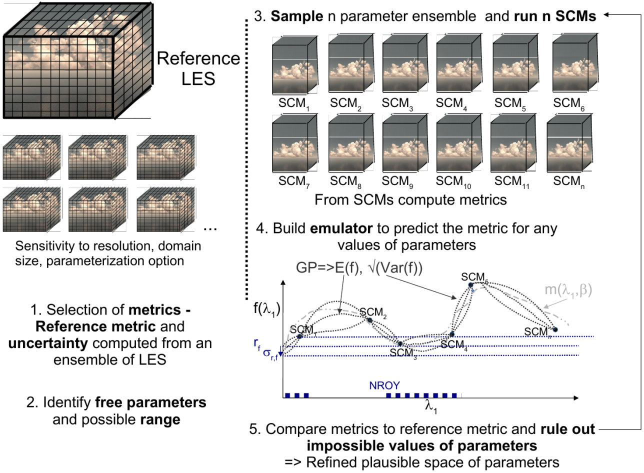
- Structure is given by $\mathcal{P}$, we are trying to calibrate values of a vector of parameters λ
- Multiple metrics we wish to satisfy. For each metric f, define a distance given by:

$$I_f(\lambda) = \frac{\| r_f - E_f[\lambda] \|}{\sigma^2_{r,f} + \sigma^2_{d,f} + Var[f(\lambda)]}$$

- Euclidean distance over history normalized by error (observational, structural, chaotic)
- Sample λ space as exhaustively as practical for $I < T$, the NROY space. Iterate in *waves*. Can use different metrics in subsequent waves.

$$\mathrm{NROY}^n = \cap_k \mathrm{NROY}_{f_k}$$

Couvreux et al (2020)



Sensitivity to resolution, domain size, parameterization option

1. Selection of **metrics - Reference metric** and **uncertainty** computed from an ensemble of LES

2. Identify **free parameters** and possible **range**

3. **Sample** n parameter ensemble and **run n SCMs**

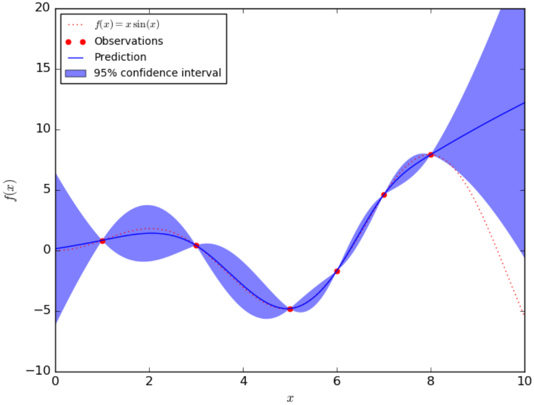
From SCMs compute metrics

4. Build **emulator** to predict the metric for any values of parameters

5. Compare metrics to reference metric and **rule out impossible values of parameters**
=> Refined plausible space of parameters

- LES as ground truth, multiple variants to get “observational error”.
- Emulate LES using SCMs encoding all the $\mathcal{P}$.
- Latin hypercube sampling of λ
- Fit Gaussian processes to SCMs to densely sample all values of λ

Gaussian processes

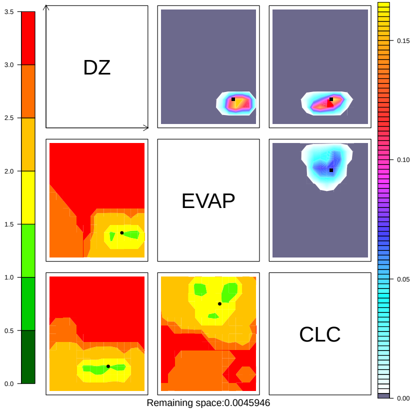


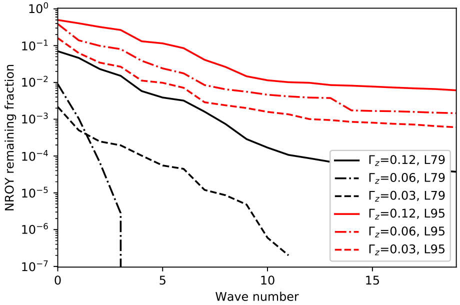
- Extremely standard emulator, widely available in python libraries
- Very poor at extrapolation, so training data must span phase space!

Couvreux et al 2020, some highlights

- Importance of a library of distinct physical regimes (e.g marine, continental) sampled by LES
- Results are sensitive to LES turbulence closure and numerics.
- Don't do sensitivity analysis on the full phase space (premise is that most of it is unphysical). But see discussion of order of imposition of metrics.
- Even individual $\mathcal{P}$ may have multiple tunable subsystems with compensating errors, e.g EDMF.
- Rule of thumb: need $10\times\mathrm{rank}(\lambda)$ SCM runs.

Hourdin et al (2020)





- Eliminate implausible parameter space comparing SCMs with LES.
- ... leaving irreducible (“structural”) model error.

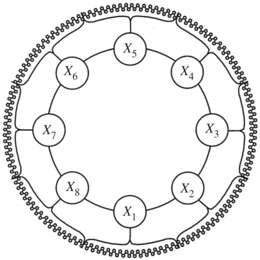
Hourdin et al 2020, highlights

- Cost of “conventional” tuning: 15 versions of the atmospheric model tuned in AMIP mode, 3-10 parameters per “wave”, 1-5 iterations of “tedious” sensitivity analysis.
- A new generation of model $\mathcal{P}_1$ is “improved” if there is one choice of λ which is better than $\mathcal{P}_0$ for *any* choice of λ.
- 9 parameters tested, 2 versions of the model, L79 and L95: L95 presumed to have smaller structural error.
- On final NROY, full tests with 3D AMIP configuration: 45 2-year runs. 1 coupled model run of 50 years (but with “fudged” ocean albedo)..

Bibliography

- Couvreux et al 2020: Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement.
- Hourdin et al 2020: Process-based climate model development harnessing machine learning: II. model calibration from single column to global
- Hourdin et al 2017: The art and science of climate model tuning.
- Williamson et al 2013: History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble
- Williamson et al 2017: Tuning without over-tuning: parametric uncertainty quantification for the NEMO ocean model

Lorenz 96, a nice abstraction

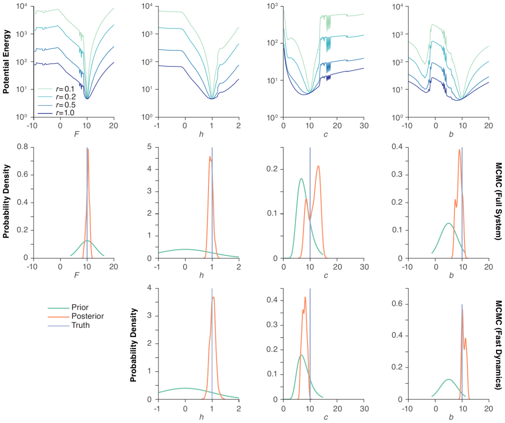


$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k} + f \tag{1}$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,l}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}X_k \tag{2}$$

- A simplified multiscale system (X and Y can stand for resolved/unresolved, slow/fast), where coupling strength can be varied... maybe too interesting? See metastability issues in Schneider et al (2017).
- Maybe too simple? (from Stephan Rasp's blog)

Lorenz96 in perfect model setting



From Schneider et al 2017. Learn Lorenz96 parameters F, h, $\log c$, b from prior “truth” run.

CliMA: Calibrate, Emulate, Sample

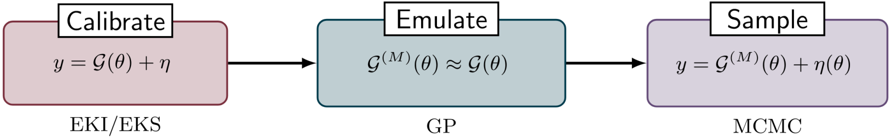

Calibrate
$y = \mathcal{G}(\theta) + \eta$
EKI/EKS
Emulate
$\mathcal{G}^{(M)}(\theta) \approx \mathcal{G}(\theta)$
GP
Sample
$y = \mathcal{G}^{(M)}(\theta) + \eta(\theta)$
MCMC

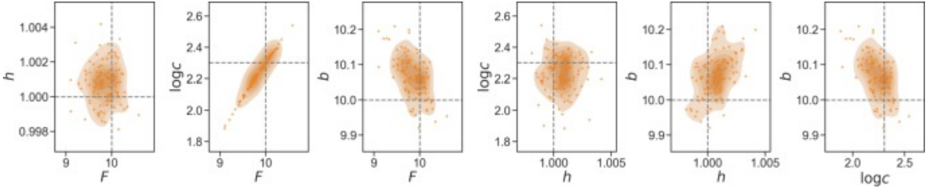
Fig. 1. Schematic of approximate Bayesian inversion method to find θ from y. EKI/EKS produce a small number of approximate (expensive) samples $\{\theta^{(m)}\}_{m=1}^{M}$. These are used to train a GP approximation $\mathcal{G}^{(M)}$ of $\mathcal{G}$, used within MCMC to produce a large number of approximate (cheap) samples $\{\theta^{(n)}\}_{n=1}^{N_s}$, $N_s \gg M$.

- *Calibrate*: approximately locate attractor using expensive forward model
- *Emulate*: cheap GP emulator to map parameter space near attractor
- *Sample*: MCMC sampling of parameter space for uncertainty quantification (parameter vector with error bounds)

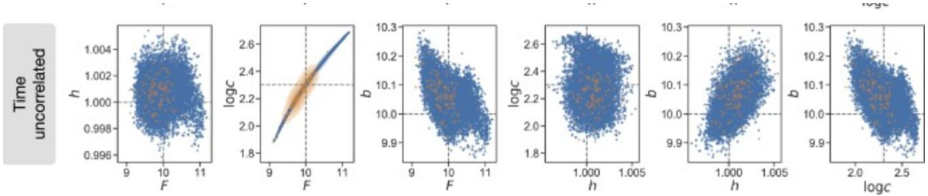
Applied to boundary layer and shallow cloud (EDMF) parameterizations, Cleary et al 2020, Dunbar et al 2021.

CES applied to Lorenz96

- Sparse sampling during calibrate phase (expensive):

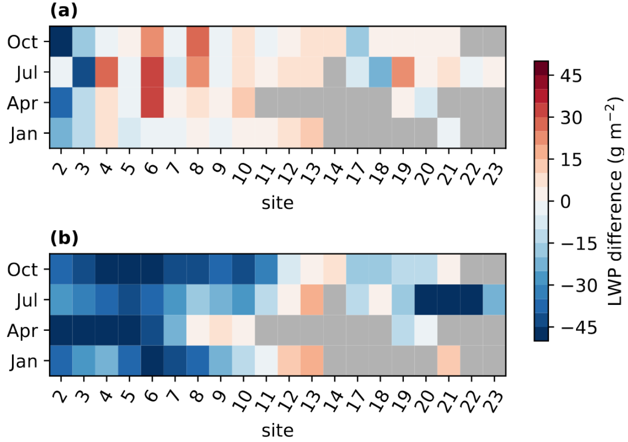


- Dense sampling using GP emulator:



From Cleary et al 2020

LES "libraries" for training

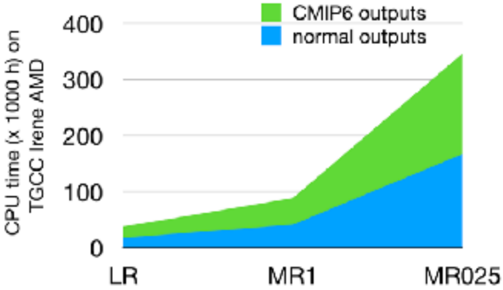


From Zhaoyi Shen et al 2021. HighTune proposes a community-based library (DEPHY).

Bibliography

- Schneider et al (2017): Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations
- Pressel et al 2017: Numerics and subgrid-scale modeling in large eddy simulations of stratocumulus clouds
- Cleary et al 2020: Calibrate, emulate, sample
- Dunbar et al 2021; Calibration and Uncertainty Quantification of Convective Parameters in an Idealized GCM
- Shen et al 2021: A Library of Large-eddy Simulations for Calibrating Cloud Parameterizations

Beyond LES: calibration of coupled models

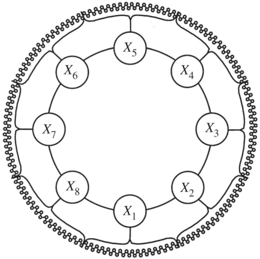


Post Hourdin et al automatic tuning:

- 5 new piCtrl coupled simulations, 250 SY each
- excessive cold biases and sea ice cover relative to baseline IPSL-CM6
- required extensive retuning of ocean and sea ice!

- GFDL experience is similar: about 50000 SY of coupled runs of CM4 and ESM4 during model calibration in addition to AMIP.

Lorenz 96 again: history matching for an “AOGCM”

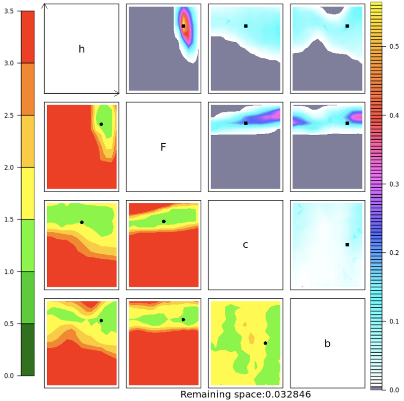


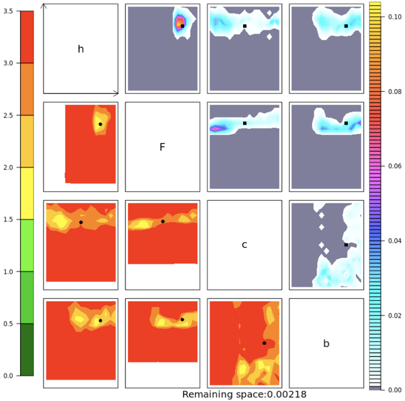
$$\frac{dX_k}{dt} = -X_{k-1}(X_{k-2} - X_{k+1}) - X_k + F - \frac{hc}{b}\sum_{j=1}^{J} Y_{j,k} + f \tag{1}$$

$$\frac{dY_{j,k}}{dt} = -cbY_{j+1,l}(Y_{j+2,k} - Y_{j-1,k}) - cY_{j,k} + \frac{hc}{b}X_k \tag{2}$$

- Similar metrics to Schneider et al (2017) $f(X, Y) = (X, \overline{Y}, X^2, X\overline{Y}, \overline{Y}^2)$
- as usual try to recover $F, h, \log c, b$ from prior “truth” run.
- AMIP: apply only Y constraints; OMIP = apply only X constraints.
- Investigate length of sample needed for training.
- Lguensat, Balaji, Deshayes 2021, *in prep.*

History Matching on Lorenz96





- History matching efficiently reduces NROY space.
- “AMIP” and “OMIP” experiments underway.
- From Lguensat, Balaji, Deshayes, *in prep.*

Outline

1. Computing technology is moving towards deep learning
2. Deep Learning for climate: no black boxes, no free lunches
3. Machine Learning for Model Calibration
   - Model calibration or “tuning”, a brief survey
   - HighTune: History matching
   - CliMA: Calibrate-Emulate-Sample
   - Calibrating coupled models with long timescales
4. Summary: prospects and challenges for model calibration using ML

To recapitulate...

- We are trying to calibrate a system with chaotic, parametric and structural uncertainty
  - Chaotic: calibrate based on time averages around an attractor.
  - Parametric: assume a closed-form expression and calibrate its parameters. Ideally with uncertainty (perhaps stochastically sampled? ... see Guillaumin and Zanna 2021), identify irreducible structural error.
  - Structural: when structural error is found, attempt to find new structures from a library of plausible terms on the RHS.
- Emulation models "up the ladder" rather than observations, for physical consistency.
- CES: scan parameter space using forward model, fill in the space using emulators, sample the space to obtain **P** and quantify uncertainty.
- HM: eliminate the implausible regions of parameter space using emulators, then sample final NROY space using forward model. There are distance metrics but no cost function: no relative weight assigned to metrics. Structural error is identified when distance above tolerance.
- Long calibration timescales are still an open problem.

Prospects and challenges

- The prospects for ML entirely replacing conventional models is very remote.
- Exact replicas (“digital twins”) may be useful for some purposes, but not all.
- ML may be a tool to derive traceable model hierarchies: *Climbing down Charney’s ladder*, Balaji (2021).
- Models must be able to go outside observational bounds, simulate counterfactuals.
- Initial focus: ML-derived emulators to sample model uncertainty, identify structural errors, quantify uncertainties.
- Multiple timescales still a problem in model calibration. Spinup as well, but that’s another story...
- M2LInES: new initiative including GFDL scientists to constrain coupled models using ML and data assimilation.
- Playing around with Lorenz96 is all very well, but to get beyond toy models the big labs (GFDL, I’m looking at you...) must get in the game.
- MLINT, the CIMES-GFDL ML interest group: journal club, expertise exchange, launch new initiatives, new cross-divisional and CIMES-GFDL collaborations.