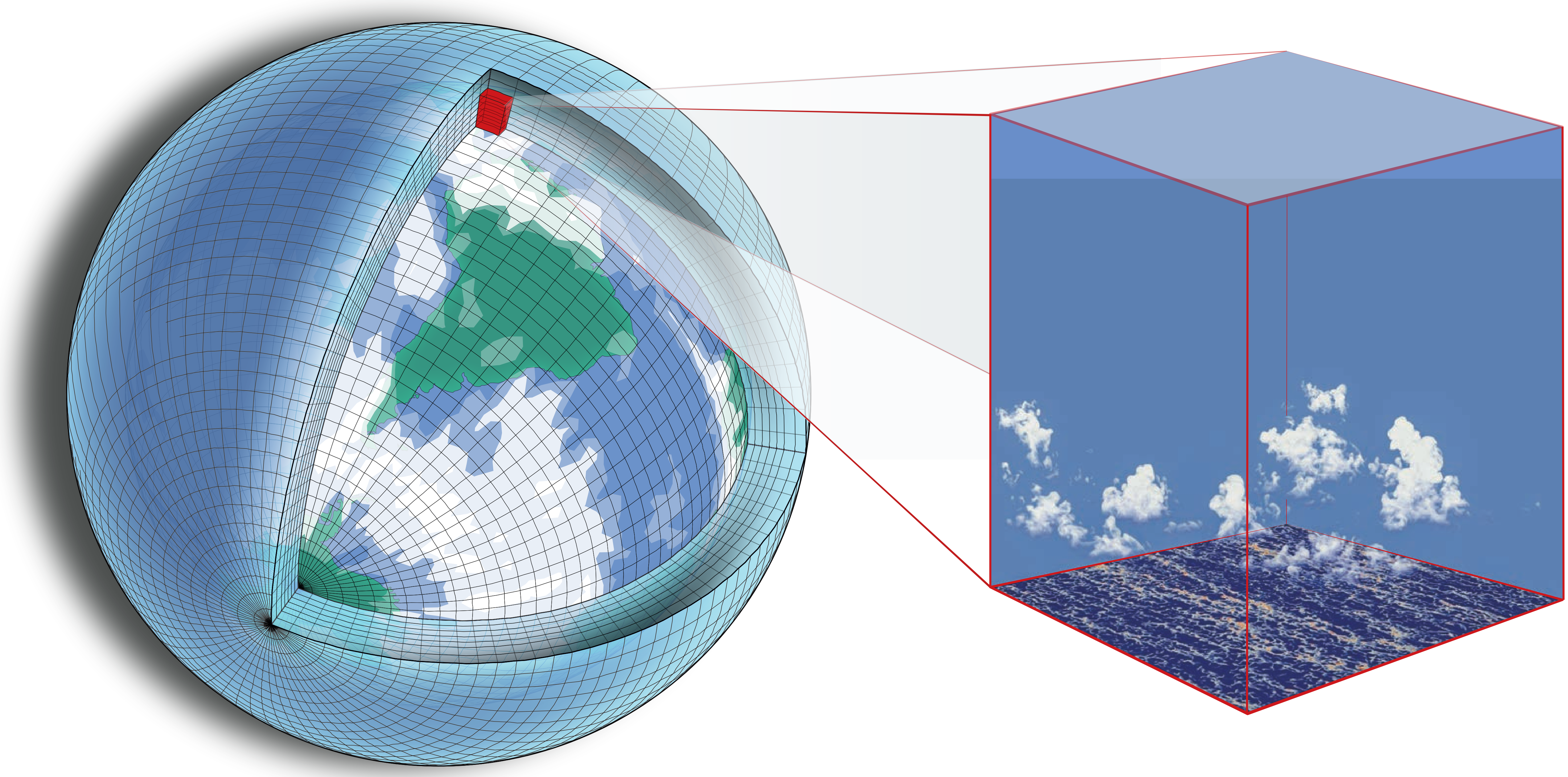
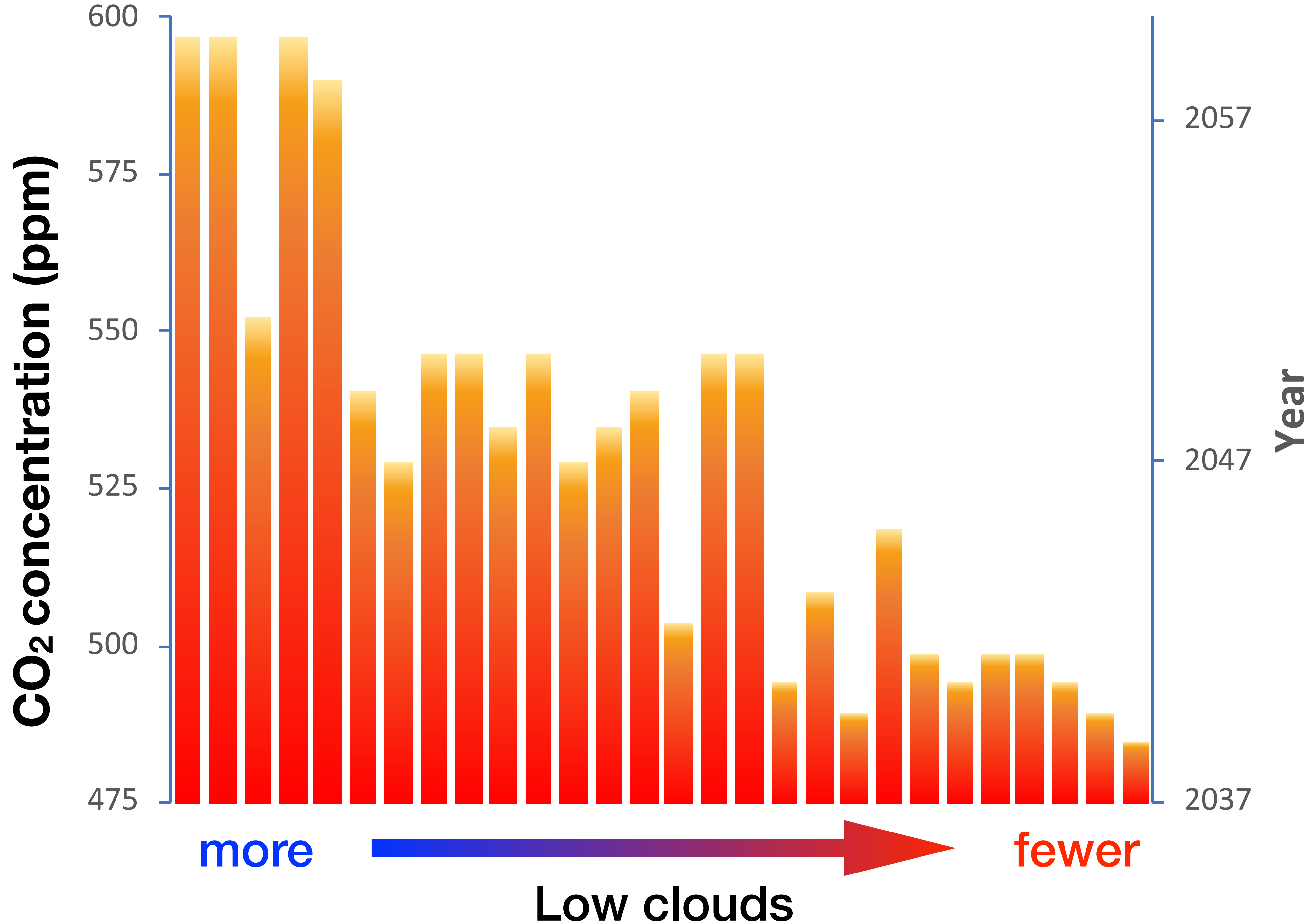


Data-Informed Climate Models With Quantified Uncertainties

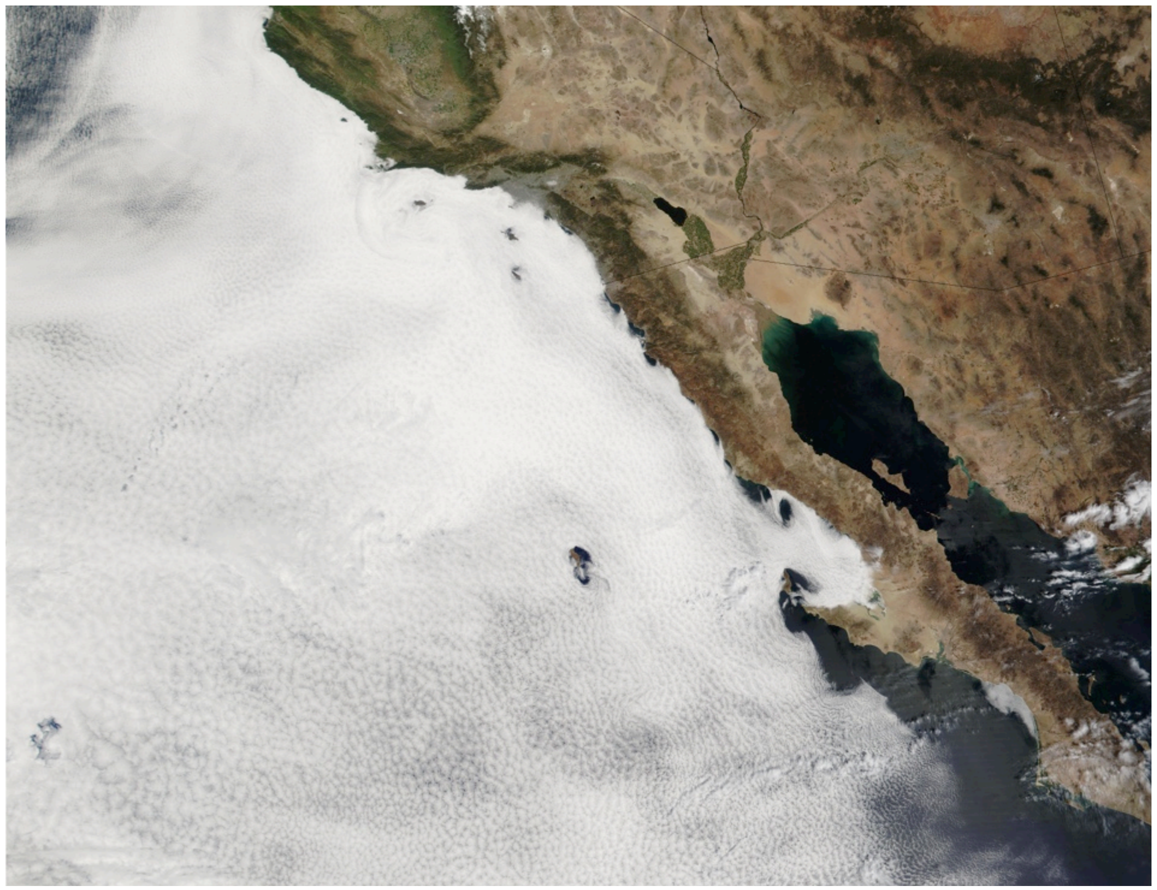
Tapio Schneider, Andrew Stuart (and the CliMA Team, especially Yair Cohen, Anna Jaruga, Jia He, Ignacio Lopez-Gomez, Zhaoyi Shen, Emmet Cleary, Ollie Dunbar, and Alfredo Garbuno)

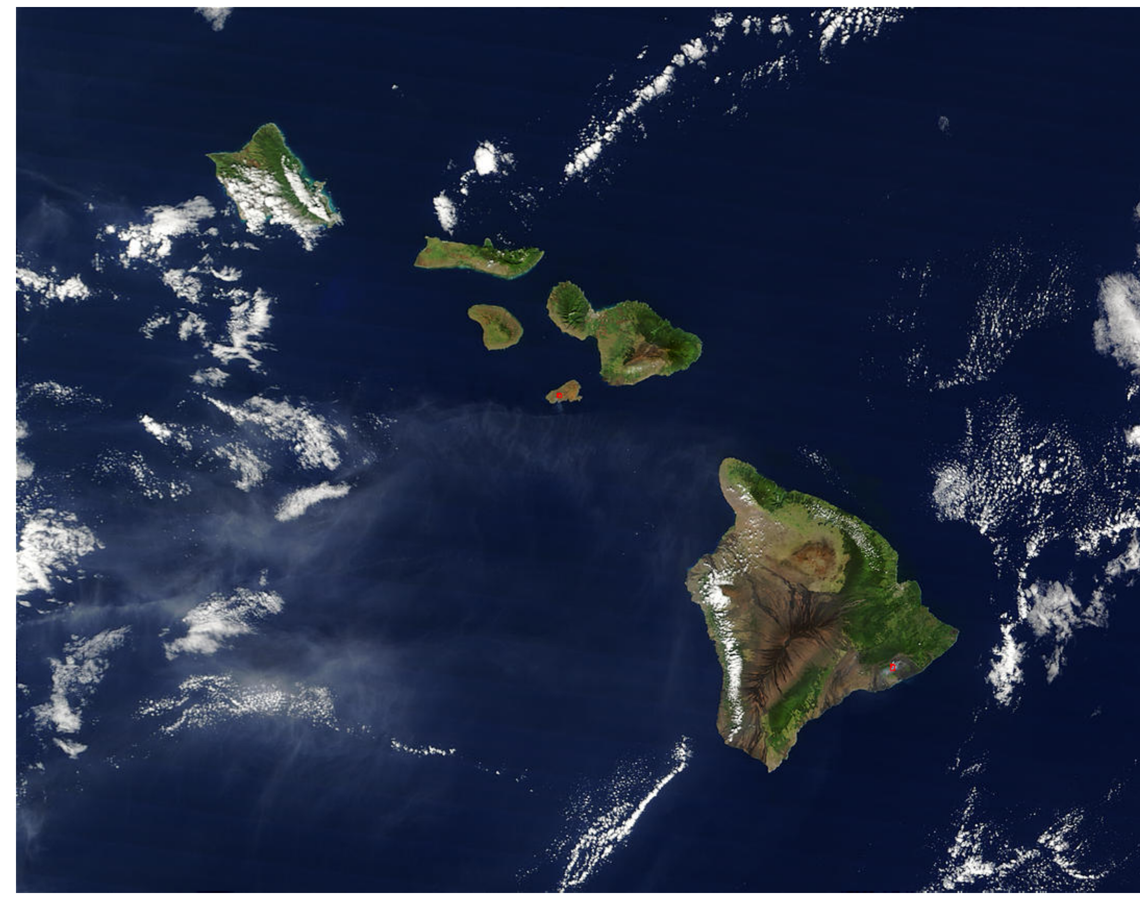
Spread in predictions for next ~30-50 years is dominated by uncertainties in low clouds; uncertainties are poorly quantified



Schneider et al., *Nature Climate Change* 2017

The primary (but not only) source of uncertainties in climate predictions is the representation of low clouds in models





http://eoimages.gsfc.nasa.gov

Stratocumulus: colder

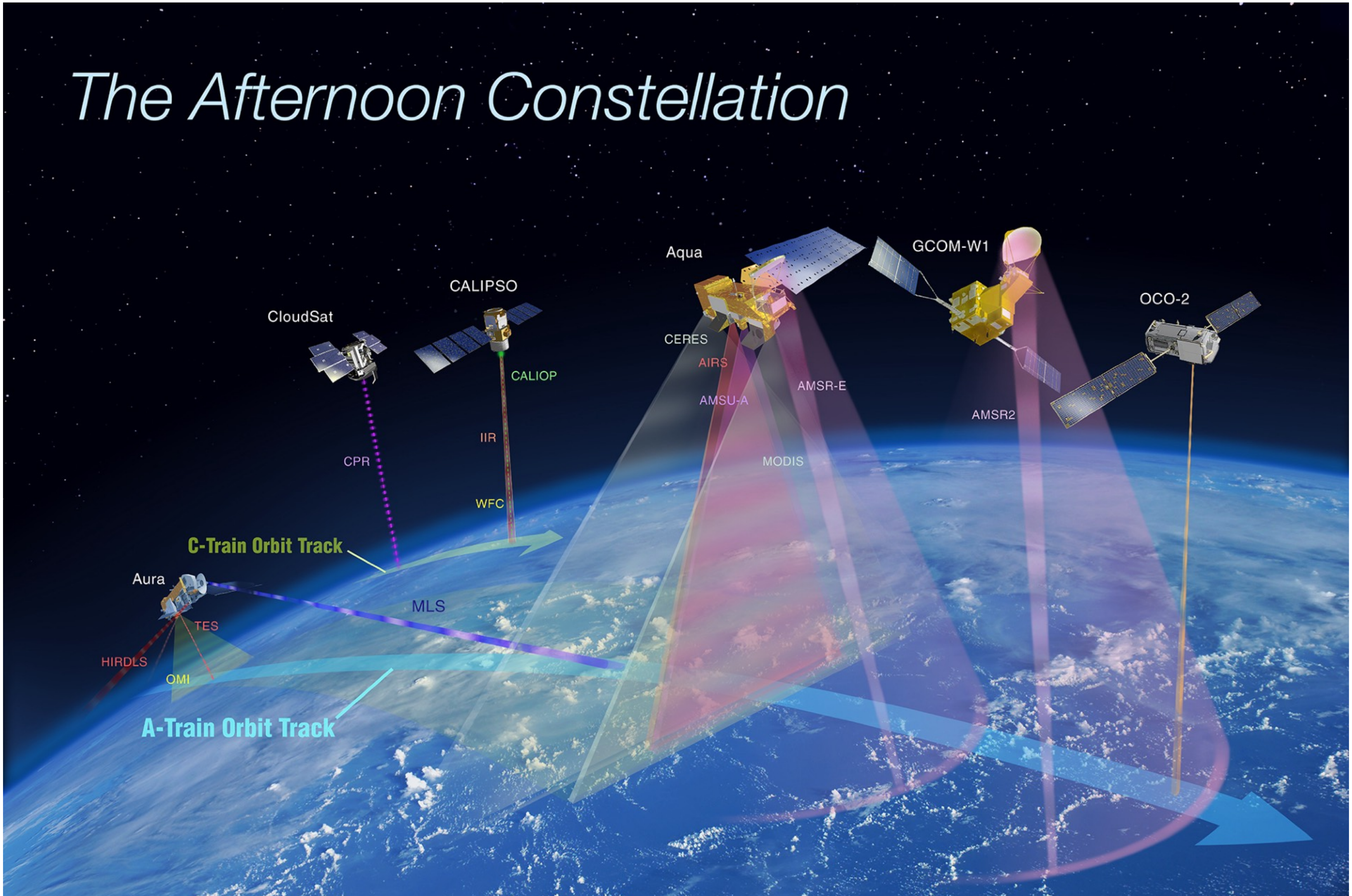
Cumulus: warmer

We don't know if we will get more low clouds (damped global warming), or fewer low clouds (amplified warming) with rising CO_2 levels

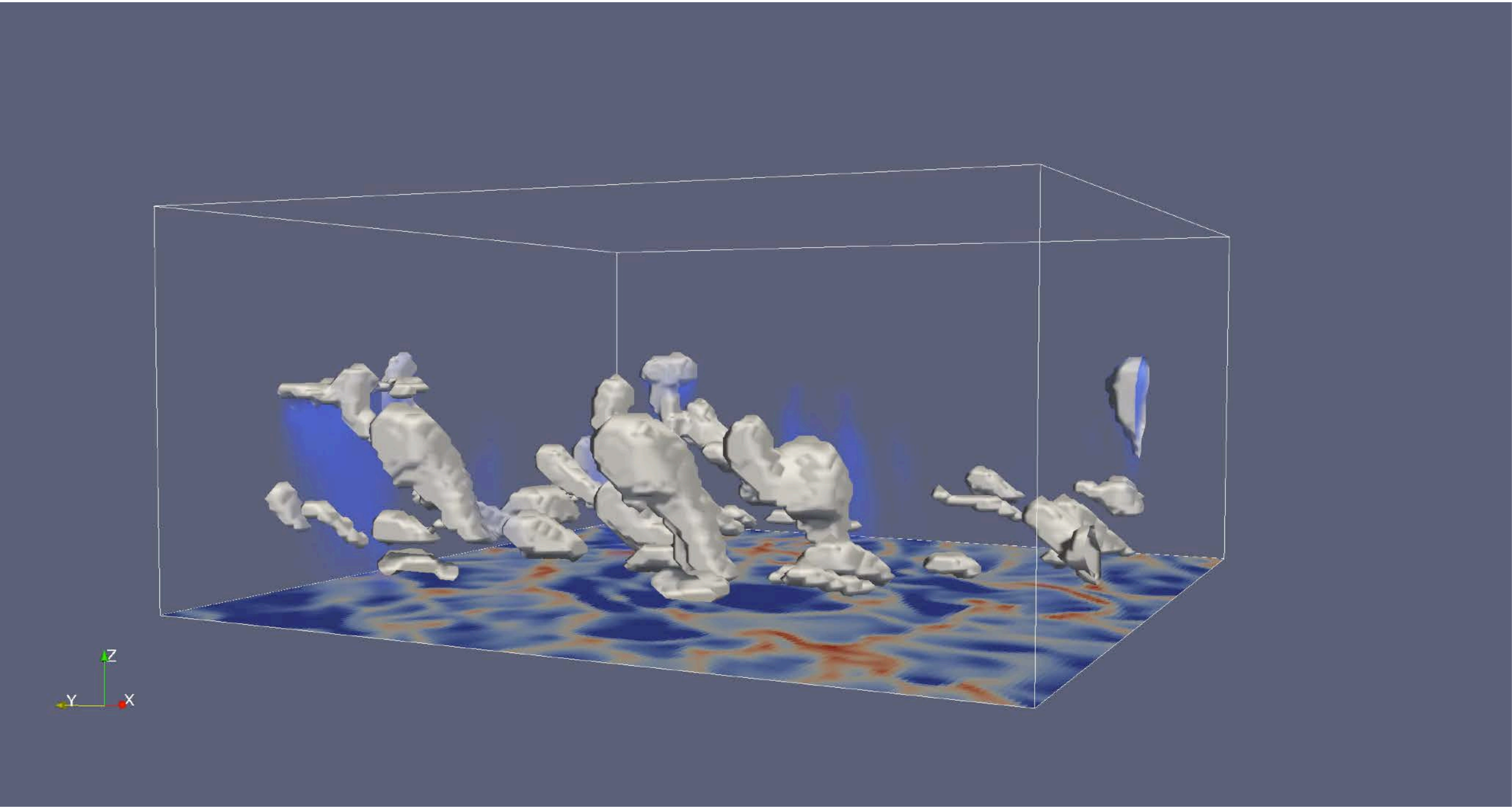
Improving predictions is urgent.

How can we make progress?

We have a wealth of global climate data, whose potential to improve models has not been tapped



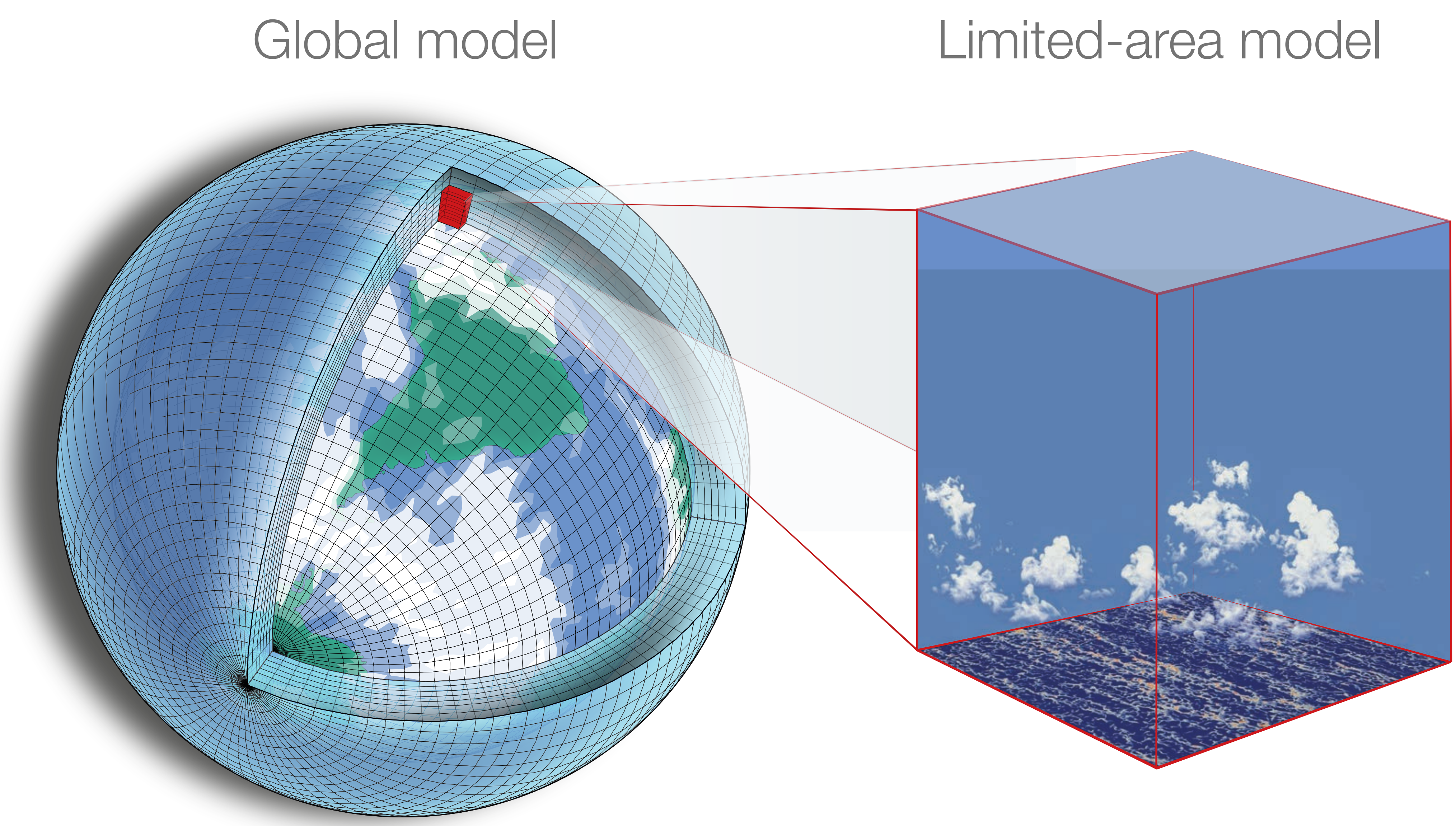
We can also simulate some small-scale processes (e.g., clouds) faithfully, albeit only in limited areas



Large-eddy simulation of tropical cumulus

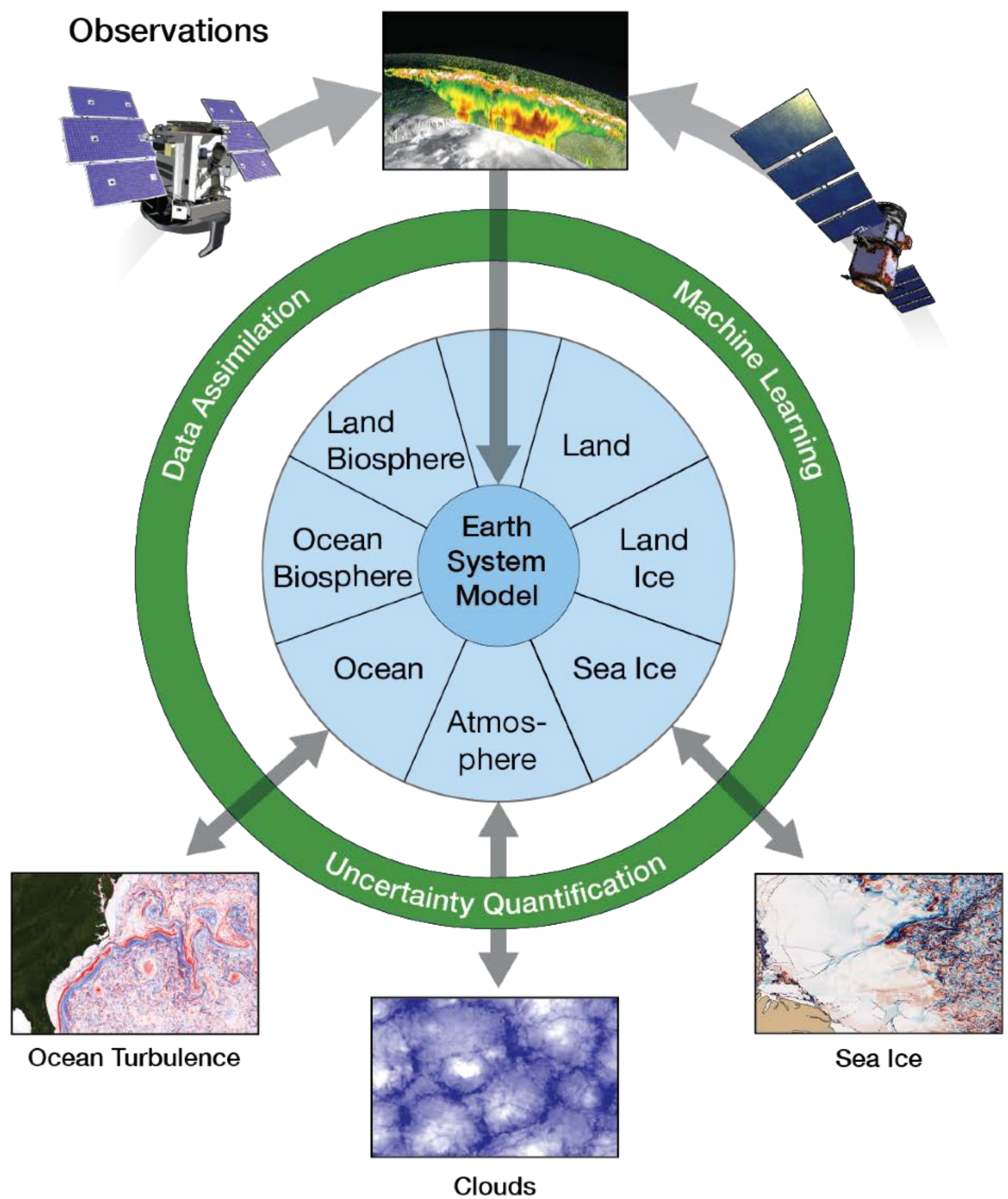
Simulation with PyCLES (Pressel et al. 2015)

Such limited area models can be nested in a global model and can, in turn, inform the global model



Thousands of high-resolution simulations can be embedded in global model in a distributed computing environment (cloud), and the global model can learn from them

Vision: build a model that learns automatically both from observations and targeted high-resolution simulations



Targeted High-Resolution Simulations

Data assimilation/ML-accelerated science

- Deep learning's success rests on overparameterization
  - Data-hungry methods
  - Leads to challenges with generalization, interpretability, and uncertainty quantification
- Success of reductionist science rests on sparsity
  - Generalizable and interpretable

Our approach: Combine both, traditional reductionist science with data science tools where reductionism reaches its limits

We want to use observations, yet need out-of-sample predictive capabilities and computational feasibility

- We need out-of-sample predictive capabilities (predict a climate we have not seen), yet want to use present-day observations
  - Use known equations of motion to the extent possible to **minimize number of adjustable parameters** and **avoid overfitting**
- Climate data often do not have high temporal resolution but do provide informative time aggregate statistics
  - Learn from **climate statistics** (in contrast to weather states in NWP)
- Running climate models is computationally extremely expensive
  - Need **fast algorithms** for learning about models from data (with judicious use of ML tools)

How does that actually work?

An example from modeling clouds.

Cloud/boundary layer turbulence schemes in current GCMs have unphysical discontinuities and many correlated parameters

- **Deep convection**: Often mass flux schemes (e.g., Arakawa & Schubert 1974, Tiedtke 1989; Arakawa & Wu 2013)
- **Shallow convection**: Often also mass flux schemes, but with discontinuously different parameters (e.g., entrainment rates)
- **Boundary layer turbulence**: Often diffusive; difficult to match with cloud layer (e.g.,Troen & Mahrt 1986)

Parametric and structural discontinuities for processes with common (e.g., dry) limits; plethora of parameters

We use a unified, physics-based model, derived by conditional averaging of equations of motion

Decomposes domain into environment (i=0) and coherent plumes (i=1, …, N):

- Continuity:

$$\frac{\partial(\rho a_i)}{\partial t} + \frac{\partial(\rho a_i \overline{w}_i)}{\partial z} + \nabla_h \cdot (\rho a_i \langle \mathbf{u}_h \rangle) = \boxed{\underbrace{\rho a_i \overline{w}_i \left(\sum_j \epsilon_{ij} - \delta_i \right)}_{\text{Mass entrainment/detrainment}}}$$

Closure functions

- Scalar mean:

$$\frac{\partial(\rho a_i \overline{\phi}_i)}{\partial t} + \frac{\partial(\rho a_i \overline{w}_i \overline{\phi}_i)}{\partial z} + \nabla_h \cdot (\rho a_i \langle \mathbf{u}_h \rangle \overline{\phi}_i) = \boxed{\underbrace{-\frac{\partial(\rho a_i \overline{w_i' \phi_i'})}{\partial z}}_{\text{Turbulent transport}}} + \boxed{\underbrace{\rho a_i \overline{w}_i \left(\sum_j \epsilon_{ij} \overline{\phi}_j - \delta_i \overline{\phi}_i \right)}_{\text{Entrainment/detrainment}}} + \underbrace{\rho a_i \overline{S}_{\phi,i}}_{\text{Sources/sinks}}$$

- Scalar covariance

$$\frac{\partial(\rho a_i \overline{\phi_i' \psi_i'})}{\partial t} + \frac{\partial(\rho a_i \overline{w}_i \overline{\phi_i' \psi_i'})}{\partial z} + \nabla_h \cdot (\rho a_i \langle \mathbf{u}_h \rangle \overline{\phi_i' \psi_i'}) = \underbrace{-\rho a_i \overline{w_i' \psi_i'} \frac{\partial \overline{\phi}_i}{\partial z} - \rho a_i \overline{w_i' \phi_i'} \frac{\partial \overline{\psi}_i}{\partial z}}_{\text{Generation/destruction by cross-gradient flux}}$$

$$+ \underbrace{\rho a_i \overline{w}_i \left[\sum_j \epsilon_{ij} \left(\overline{\phi_j' \psi_j'} + (\overline{\phi}_j - \overline{\phi}_i)(\overline{\psi}_j - \overline{\psi}_i) \right) - \delta_i \overline{\phi_i' \psi_i'} \right]}_{\text{Covariance entrainment/detrainment}} - \underbrace{\frac{\partial(\rho a_i \overline{w_i' \phi_i' \psi_i'})}{\partial z}}_{\text{Turbulent transport}} + \underbrace{\rho a_i (\overline{S_{\phi,i}' \psi_i'} + \overline{S_{\psi,i}' \phi_i'})}_{\text{Sources/sinks}}.$$

(Tan et al., JAMES 2018, Cohen et al. JAMES 2020, Lopez-Gomez et al. JAMES 2020))

Parametric functions requiring closure appear in the coarse-grained equations; can be refined with data

- Entrainment and detrainment (exchange between subdomains):

  Represented by a physical entrainment length ($|b|/w^2$) and an adjustable function of nondimensional parameters

  $$\varepsilon, \delta = c_\varepsilon \frac{1}{L} f\left(RH \ldots\right)$$

- Nonhydrostatic pressure gradients

  Represented by a combination of buoyancy reduction (virtual mass) and pressure drag

  $$-\frac{\partial p_{nh}}{\partial z} = -\rho a \left(\alpha_b \bar{b} + \alpha_d \frac{\left(\bar{w}^{up} - \bar{w}^{env}\right)\left|\bar{w}^{up} - \bar{w}^{env}\right|}{H a^{1/2}} \right)$$

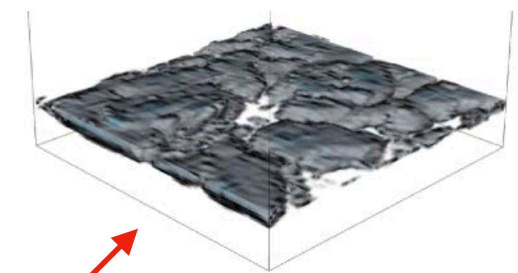
- Eddy diffusion/mixing length

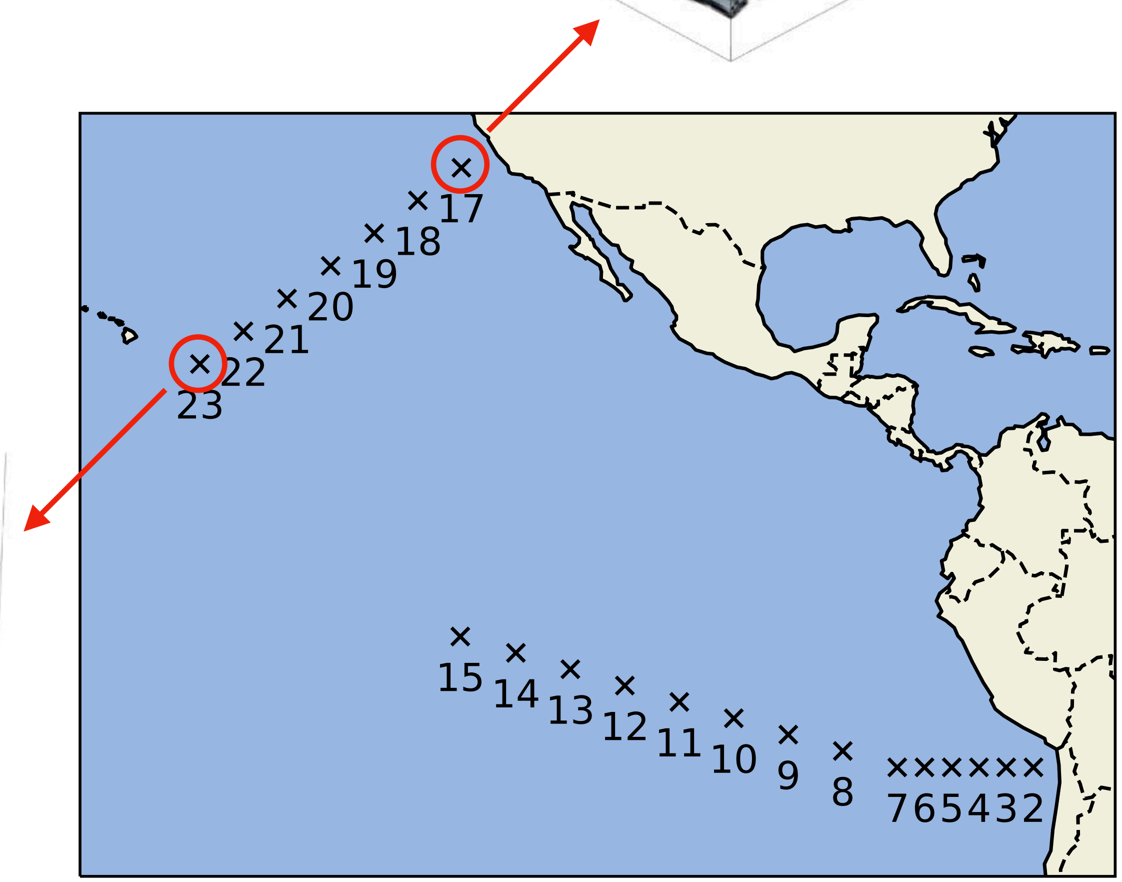
  Mixing length as soft minimum of all possible balances between production and dissipation of TKE

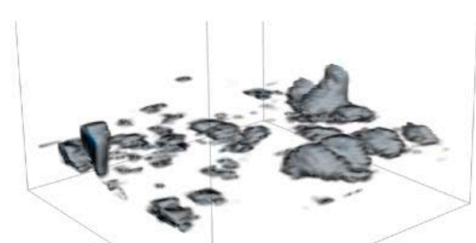
  $$K = c_k l \sqrt{TKE}$$

Calibration with suite of LES driven by GCM

- 5-year averaged monthly mean forcing from HadGEM2-A amip experiments
- Prescribed SST, RRTM, one-moment microphysics based on Kessler
- Domain size: 6km x 6km x 4km, resolution: 75m x 75m x 20m
- Simulation time: 6 days

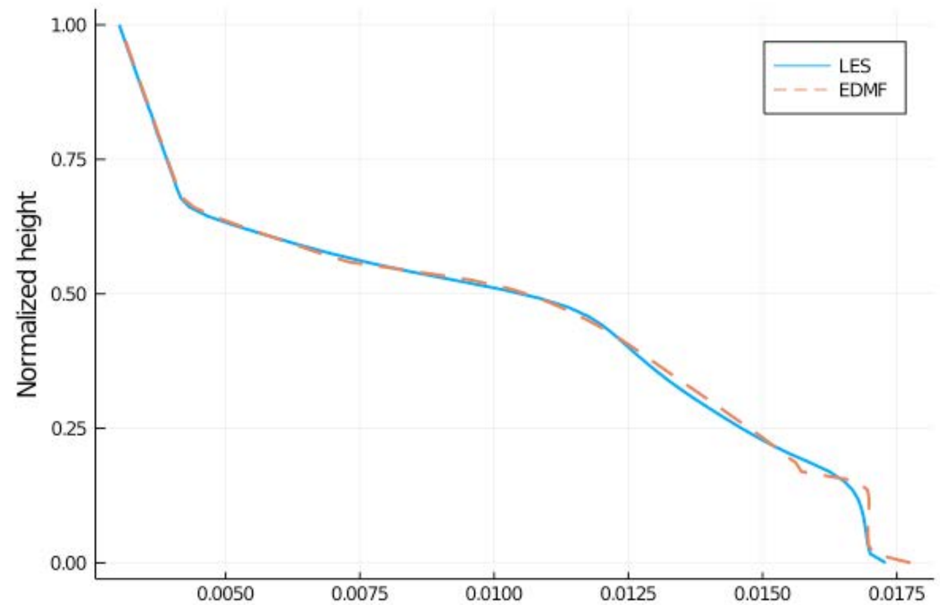




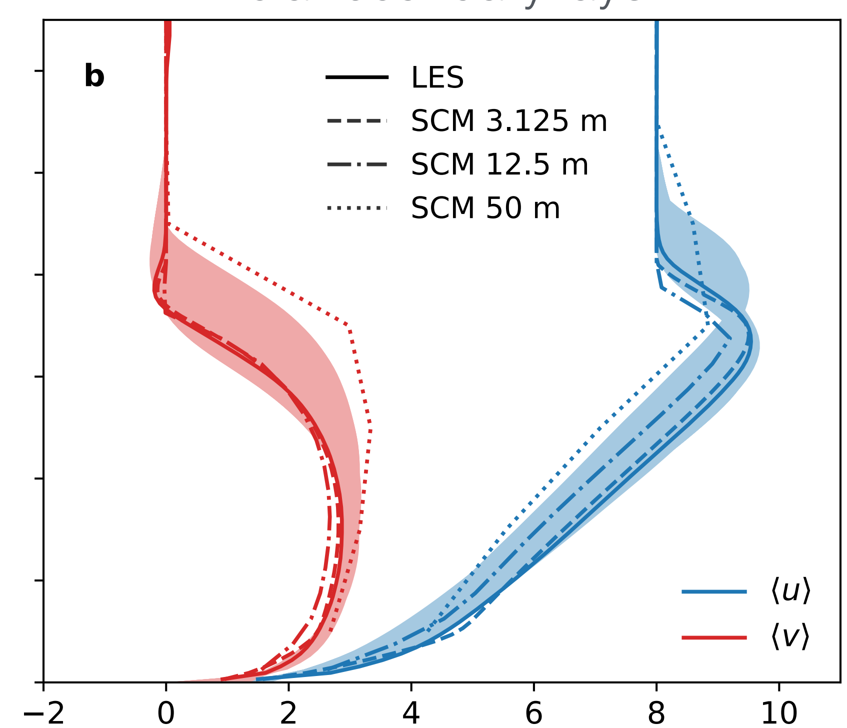


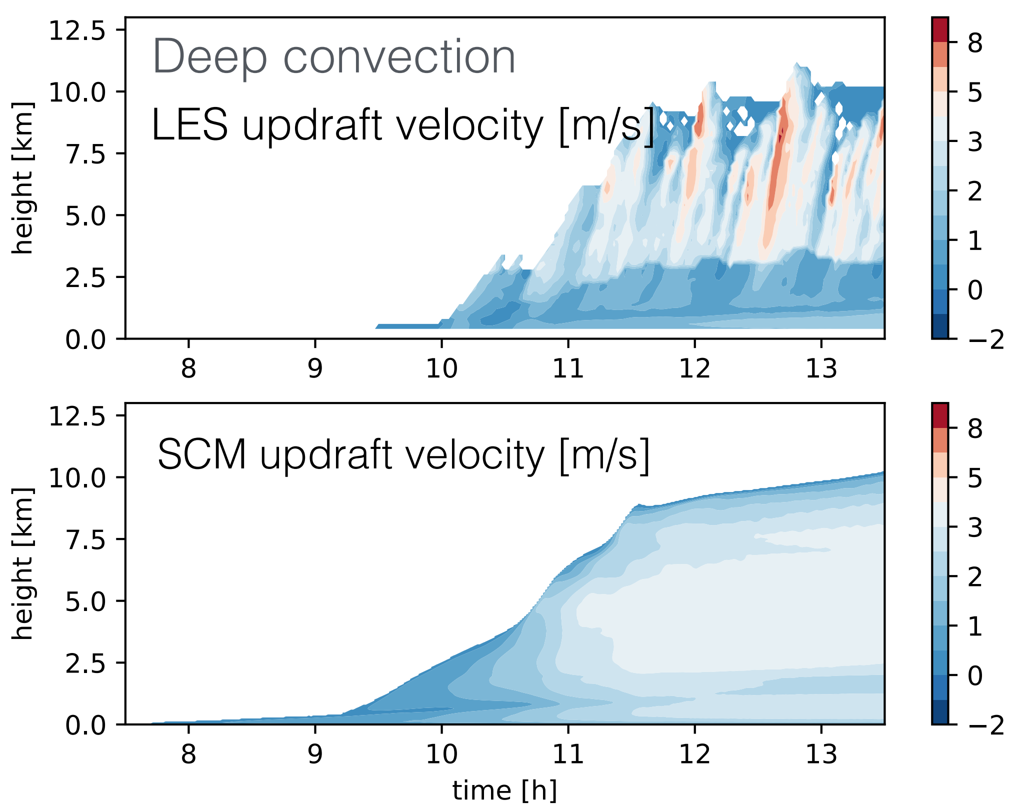
Reduced-order model captures polar and subtropical boundary layer and clouds (which have vexed climate models for decades)

Shallow cumulus

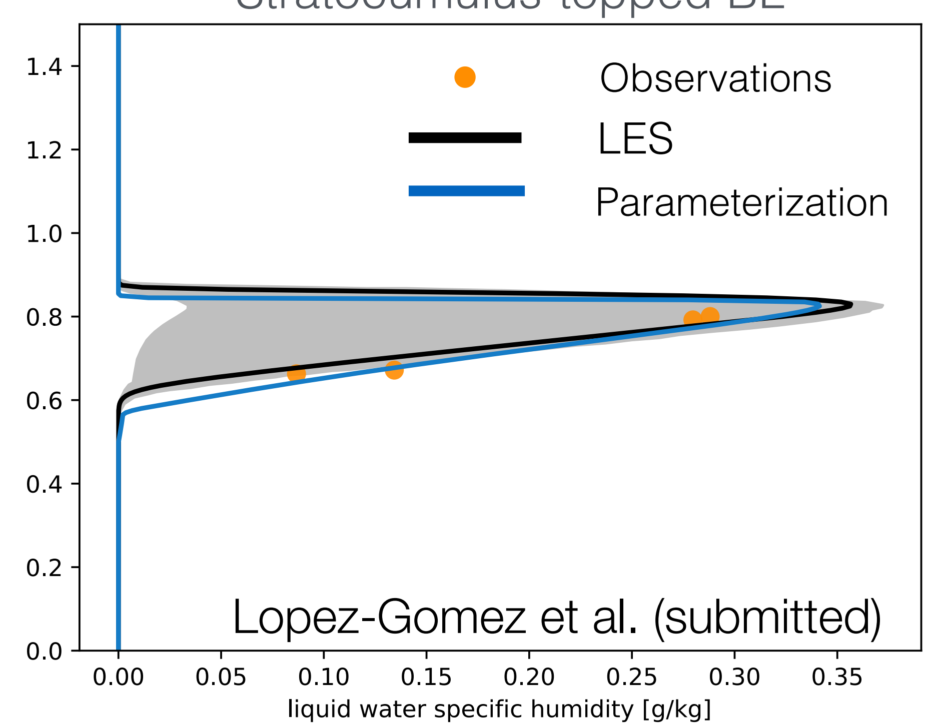


Polar boundary layer





Stratocumulus-topped BL



Calibrating a climate model and quantifying its uncertainties

Andrew Stuart

Oliver Dunbar

Alfredo Garbuno

We want to improve climate models in a similar way that weather forecasts have improved, though data assimilation approaches

We are using **statistics accumulated in time** (e.g., over seasons) to calibrate model components jointly by:

1. ***Minimizing model biases,*** *especially biases that are known to correlate with the climate response of models. That is, we will minimize mismatches between time averages of ESM-simulated quantities and data, directly targeting quantities relevant for climate predictions.*

2. ***Minimizing model-data mismatches in higher-order Earth system statistics,*** e.g., covariances such as cloud-cover/surface temperature covariances, which are known to correlate with the climate response of models. Higher-order statistics relevant for predictions (e.g., precipitation extremes) are also included in objective function.

We combine calibration and Bayesian approaches in a three step process for fast Bayesian learning

Calibrate → **Emulate** → **Sample**

Calibrate

$$y = \mathcal{G}(\theta) + \eta$$

Sample parameter space efficiently (EKS/EKI)

Emulate

$$\mathcal{G}^{(M)}(\theta) \approx \mathcal{G}(\theta)$$

Train emulator on calibration data

Sample

$$y = \mathcal{G}^{(M)}(\theta) + \eta(\theta)$$

MCMC sampling from emulator to get posterior density

- Experimental design (where to place high-resolution simulations) can be incorporated into CES pipeline
- Gives approximate Bayesian posterior (i.e., quantified uncertainties, including covariance structure of error etc.)

Cleary, Garbuno-Inigo, Lan, Schneider, Stuart, J. Comp. Phys., 2020

Proof-of-concept in idealized general circulation model (GCM)

- GCM is an idealized aquaplanet model
- It has a simple convection scheme that relaxes temperature and specific humidities to reference profiles

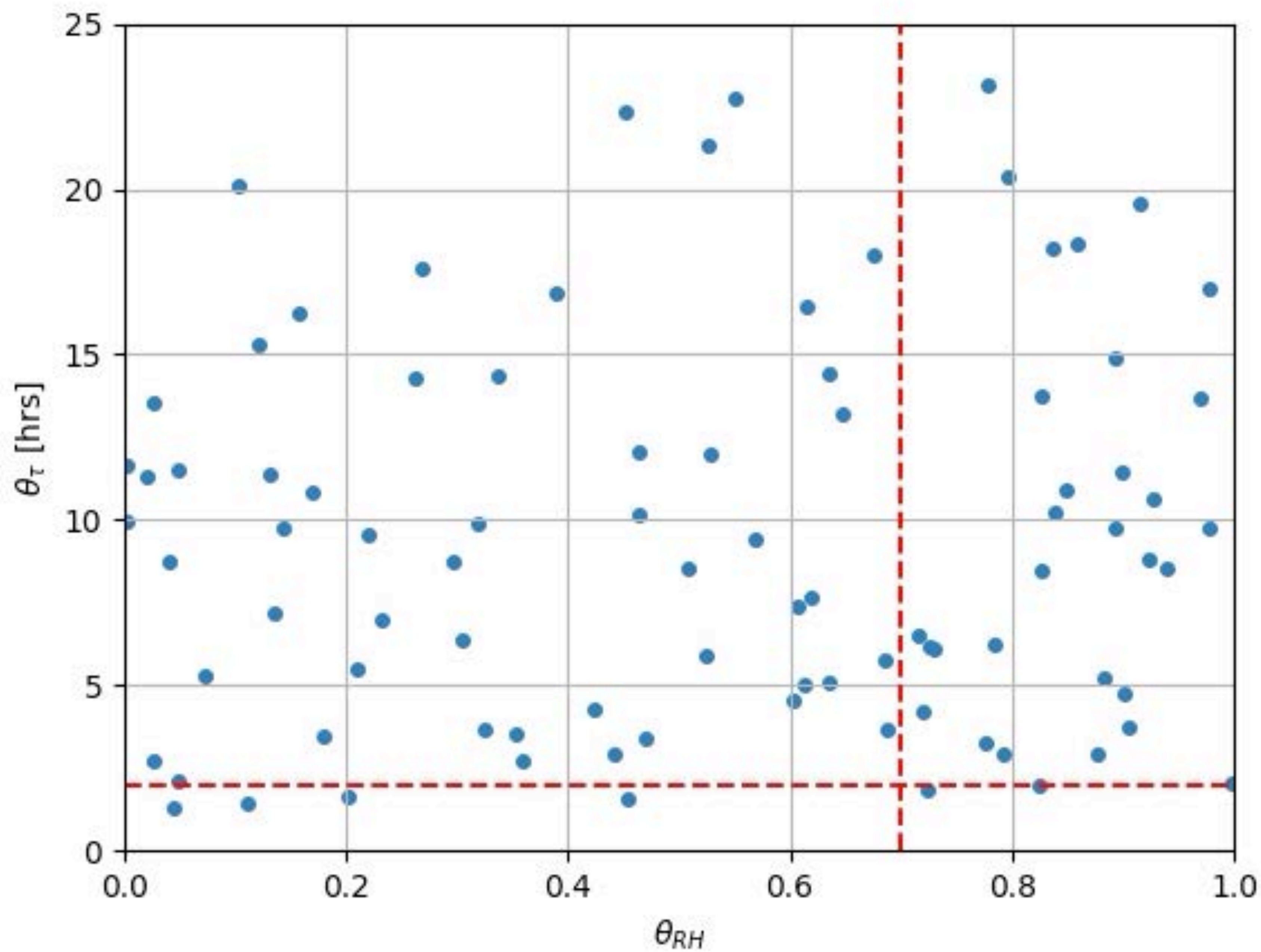
$$\partial_t T + v \cdot \nabla T + \cdots = -\frac{T - T_{\mathrm{ref}}}{\tau}$$

$$\partial_t q + v \cdot \nabla q + \cdots = -\frac{q - \mathrm{RH}_{\mathrm{ref}} q^*(T_{\mathrm{ref}})}{\tau}$$

- Two closure parameters: **timescale** τ and **reference relative humidity** $\mathbf{RH_{ref}}$

(Dunbar et al., submitted, https://arxiv.org/abs/2012.13262)

(1) **Calibrate** with ensemble Kalman inversion

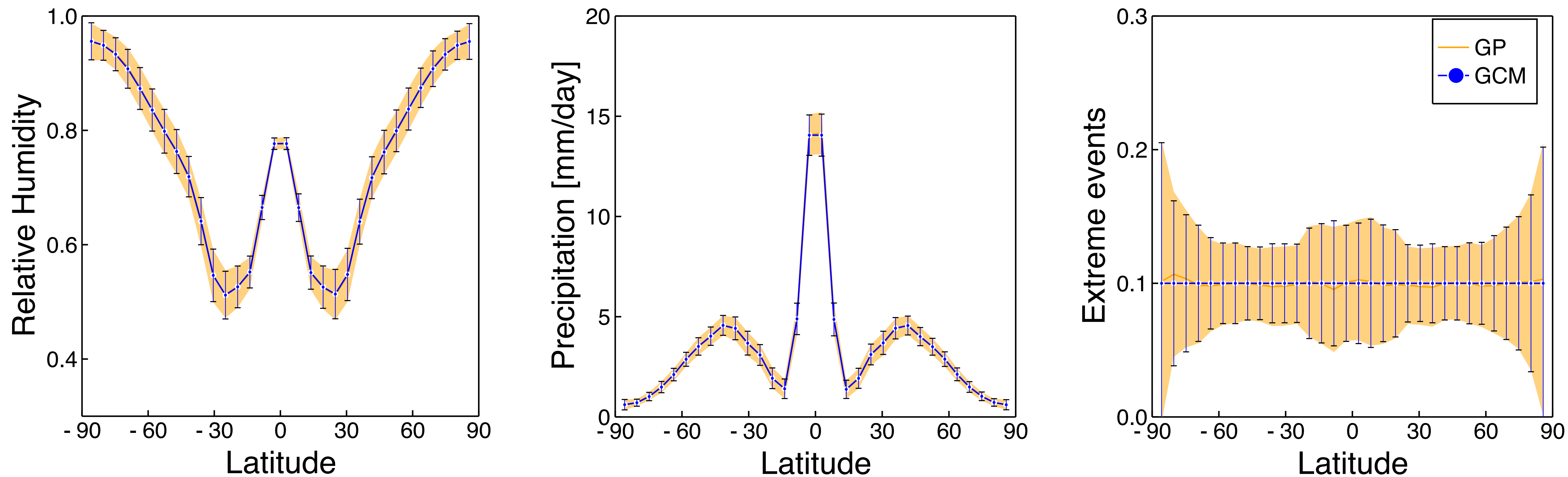


Objective function has ***relative humidity***, ***mean precipitation***, *and* ***precipitation extremes***

Ensemble Kalman inversion for parameters in convection scheme: ensemble of size 100 converges in ~5 iterations

Courtesy Emmet Cleary

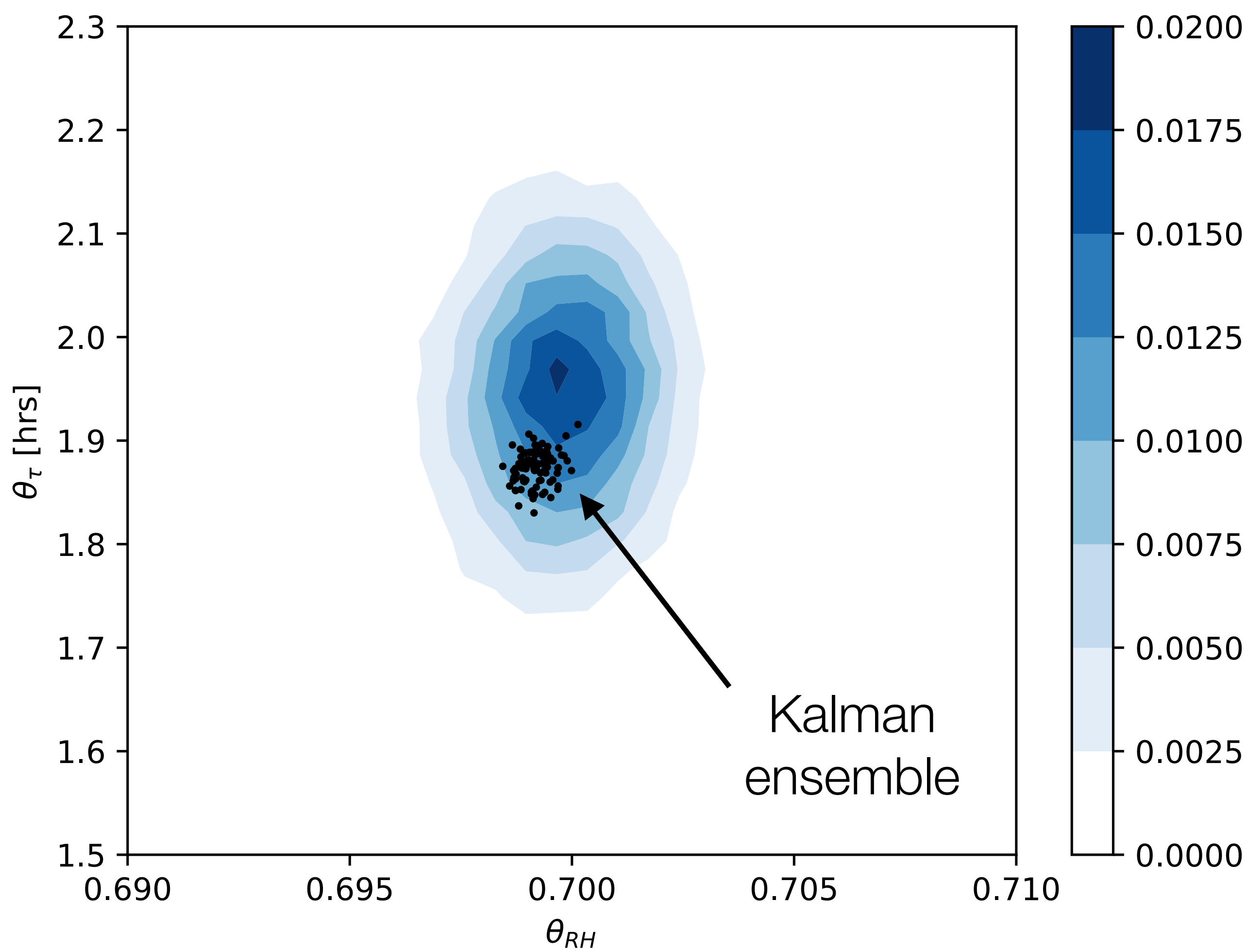
(2) **Emulate** parameters-to-statistics map during calibration step with Gaussian processes



Effective emulation of model statistics at vanishing marginal cost; additional important advantage: smoothing of objective function (can be replace by NNs for better scaling)

(3) **Sample** emulator to obtain posterior PDF for uncertainty quantification

MCMC (500,000 iterations) on GP trained on ensemble gives good estimate of posterior PDF

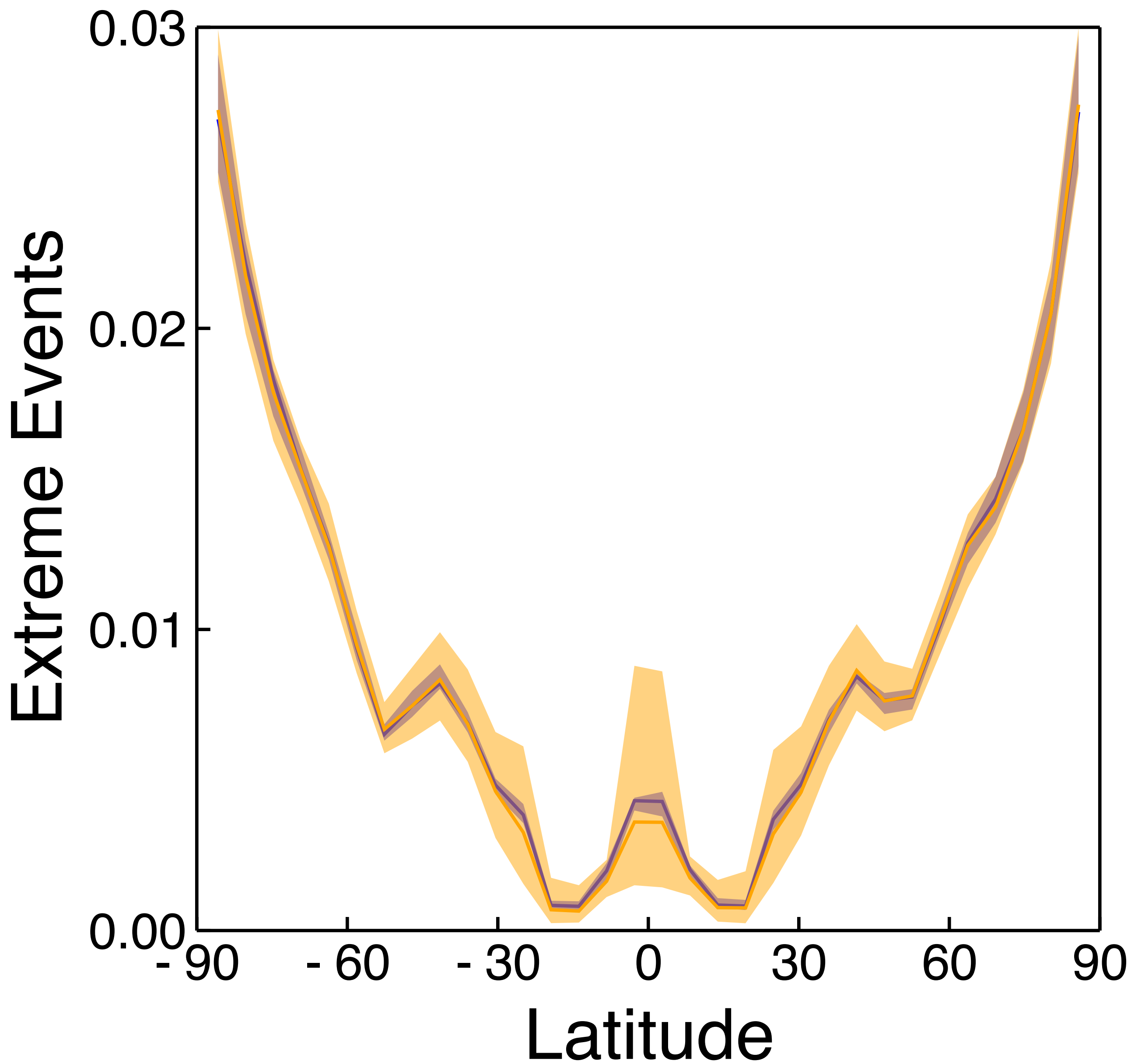


	σ_{RH}	σ_τ (hrs)
MCMC (EKI-GP)	$3.32\text{x}10^{-3}$	0.168
MCMC (Gold Standard)	$2.82\text{x}10^{-3}$	0.169

Approximate Bayesian inversion at 1/1000th the cost of standard methods
First calibrate-emulate-sample paper: https://arxiv.org/abs/2001.03689

We can then draw an ensemble of climate predictions from the posterior of parameters, for UQ of predictions

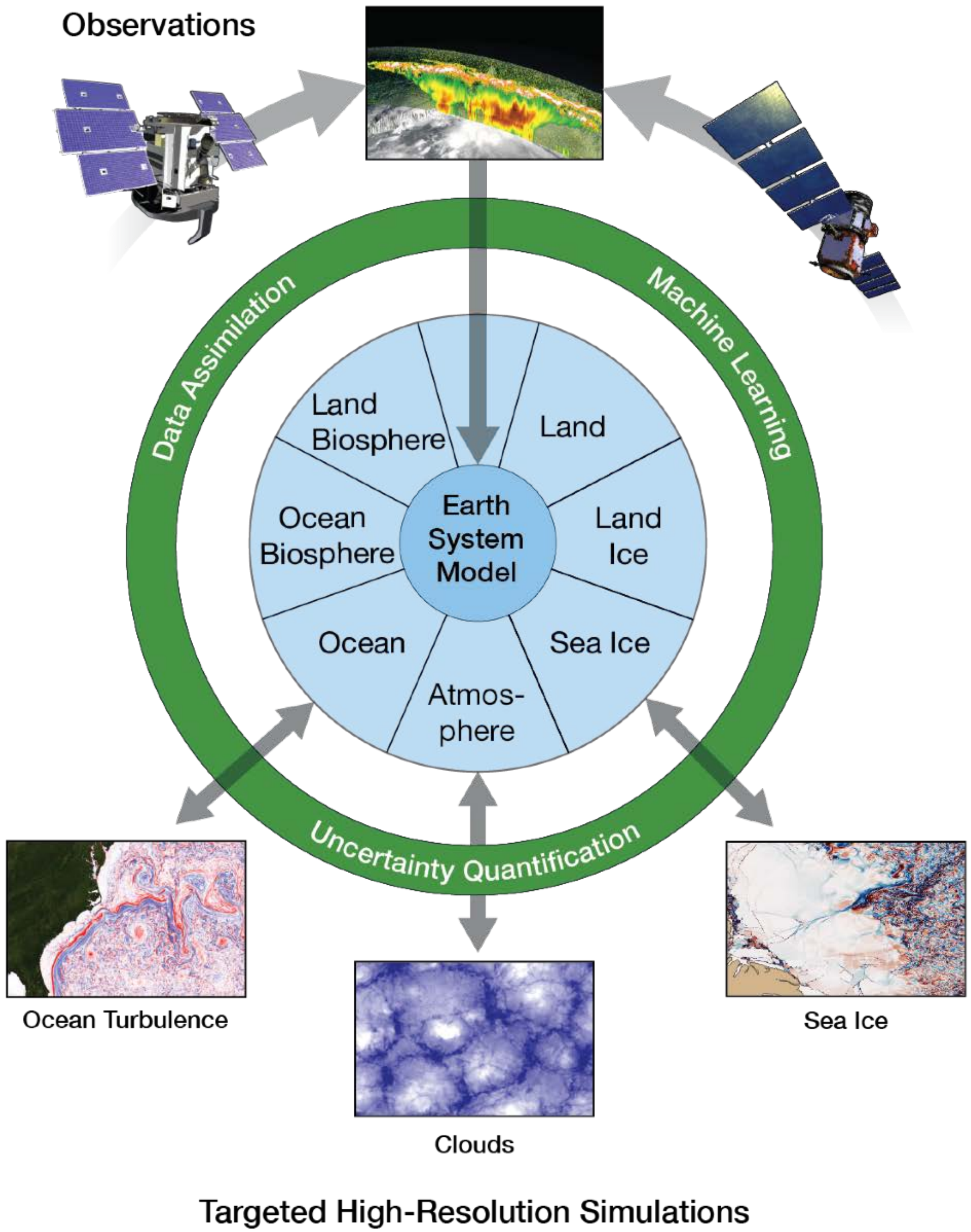
Probability of exceeding 99.9th percentile of control precipitation in warmer climate



- Quantifying risks in tails of distributions is crucial for assessing risks of climate impacts
- Can also incorporate sparse learning about structural model error (https://arxiv.org/abs/2007.06175)

(Dunbar et al., submitted, https://arxiv.org/abs/2012.13262)

We are pursuing the same approach for all components of the new Earth system model



3-year goals

- Build a model that learns automatically from observations and high-resolution simulations
- Achieve improved simulations of present climate (e.g., rainfall distribution, rainfall extremes)
- Provide predictions with UQ (including structural errors) based on observations and high-resolution simulations

Conclusions

- **Reducing and quantifying uncertainties** in climate models is urgent but within reach
- To reduce and quantify uncertainties, we **combine *process-informed* models with *data-driven* approaches using *climate statistics***
- **Physics-based subgrid-scale models** can capture turbulence and cloud regimes that have vexed climate models for decades
- Our subgrid-scale models will **learn both from observations and (where possible) from high-resolution simulations spun off on the fly**
- **Calibrate-emulate-sample** forms the core of the data assimilation/ machine learning layer and achieves up to 1,000x speed-up relative to traditional Bayesian learning methods

Much interesting work (SGS models, more effective filtering strategies, optimal targeting of high-res simulations…) remains to be done!

With thanks to CliMA's funders

ERIC AND WENDY SCHMIDT

SCHMIDT **FUTURES**

HEISING-SIMONS FOUNDATION

CHARLES TRIMBLE

RONALD AND MAXINE LINDE CLIMATE CHALLENGE

NSF