Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA

ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

From Few to Many Maps: A Fast Map-Level Emulator for Extreme Augmentation of Small Datasets

Paolo Campeti - INFN Sezione di Ferrara, ICSC
In collaboration with J.-M. Delouis, L. Pagano, E. Allys, M. Lattanzi, M. Gerbino

Spoke 3 III Technical Workshop, Perugia May 26 -29, 2025

Scientific Rationale

Massive Monte Carlos of end-to-end TOD simulations of CMB systematic effects are highly computationally expensive

- Typical **simulation campaign** $\mathcal{O}(10^3)$ maps costs $\mathcal{O}(100)$ **million CPU hrs** (e.g. *Planck*)
  - Limited number of simulations → high sample variance in empirical covariance matrices → non-optimal error bars, biased inverse covariance
- For **future surveys** (e.g. *LiteBIRD*, SO, CMB-S4):
  - *High-accuracy inference* needs $\mathcal{O}(10^{4-5})$ simulations *[Beck+'22]*
  - *Simulation-Based Inference* needs $\mathcal{O}(10^{5-6})$ simulations *[Wolz+'23]*

  } $\mathcal{O}(10^{4-6})$ millions CPU hrs!?
  - *Instrument design*: repeat for many different systematics effects, different noise models to find optimal configuration
- Computational cost might make these simply unfeasible → **finding a solution is urgent!**

Solution: Generative Modelling from Latent Representation

- A good ***Generative Model*** *(or **Emulator**)* produces new *synthetic* samples which:

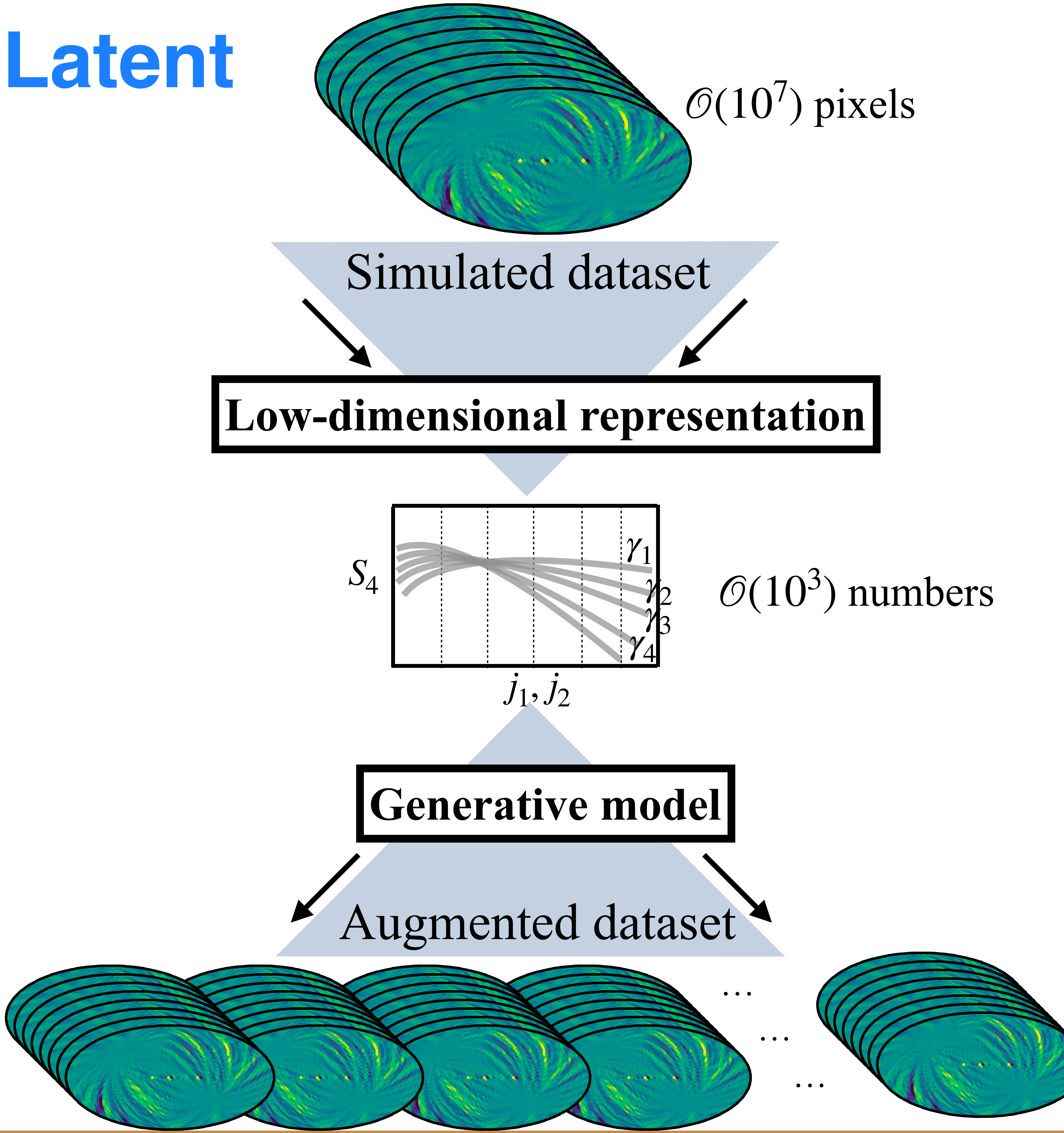
    1. reproduce true data features

    2. are representative of the true underlying data distribution

- Direct emulation often fails due to high dimensionality and/or not enough training data

- Train instead on a low-dimensional ***latent representation***!

- But GANs, VAEs, diffusion models still need massive and expensive training sets…**catch 22!**



Scattering Covariance solves the small training set problem

- Very powerful as latent representation for generative models
- Interpretable *Non-Gaussian* summary statistic inspired by CNNs *[Mallat'12, Bruna&Mallat'13, Cheng+'23]*
- (Iterative) **convolution** of field I with ***fixed* wavelet kernels** ψ_λ at oriented scale $\lambda = (j, \gamma)$ + **nonlinearity** (modulus)
- Extended also to **Cross**-Scattering Covariance between 2 maps

Convolution separates field I into individual scales

$$S_1^{\lambda_1} = \left\langle \left| I \star \psi_{\lambda_1} \right| \right\rangle$$

$$S_2^{\lambda_1} = \left\langle \left| I \star \psi_{\lambda_1} \right|^2 \right\rangle$$

$$S_3^{\lambda_1,\lambda_2} = \mathrm{Cov}\left[I \star \psi_{\lambda_1}, \left| I \star \psi_{\lambda_2} \right| \star \psi_{\lambda_1} \right]$$

$$S_4^{\lambda_1,\lambda_2,\lambda_3} = \mathrm{Cov}\left[\left| I \star \psi_{\lambda_3} \right| \star \psi_{\lambda_1}, \left| I \star \psi_{\lambda_2} \right| \star \psi_{\lambda_1} \right]$$

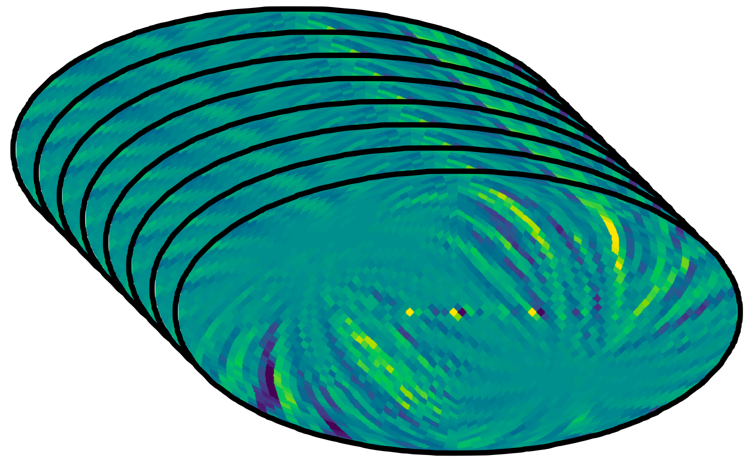
Interaction among different scales through non-linearity

Extreme Augmentation Algorithm

[Bruna & Mallat'19, Allys+'20, Price+'23, Cheng+'23, Häggbom+'24]

1. Simulation

Simulate small end-to-end *input dataset* $\{\tilde{x}_i\}$



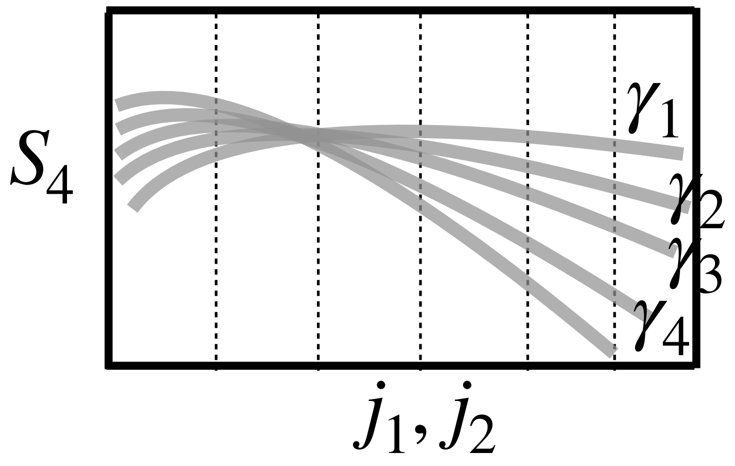
2. Latent vector of targets

Latent representation:

$$z_i = \Phi(\tilde{x}_i)$$

Scattering covariance

$$\Phi = \{S_1, S_2, S_3, S_4\}$$



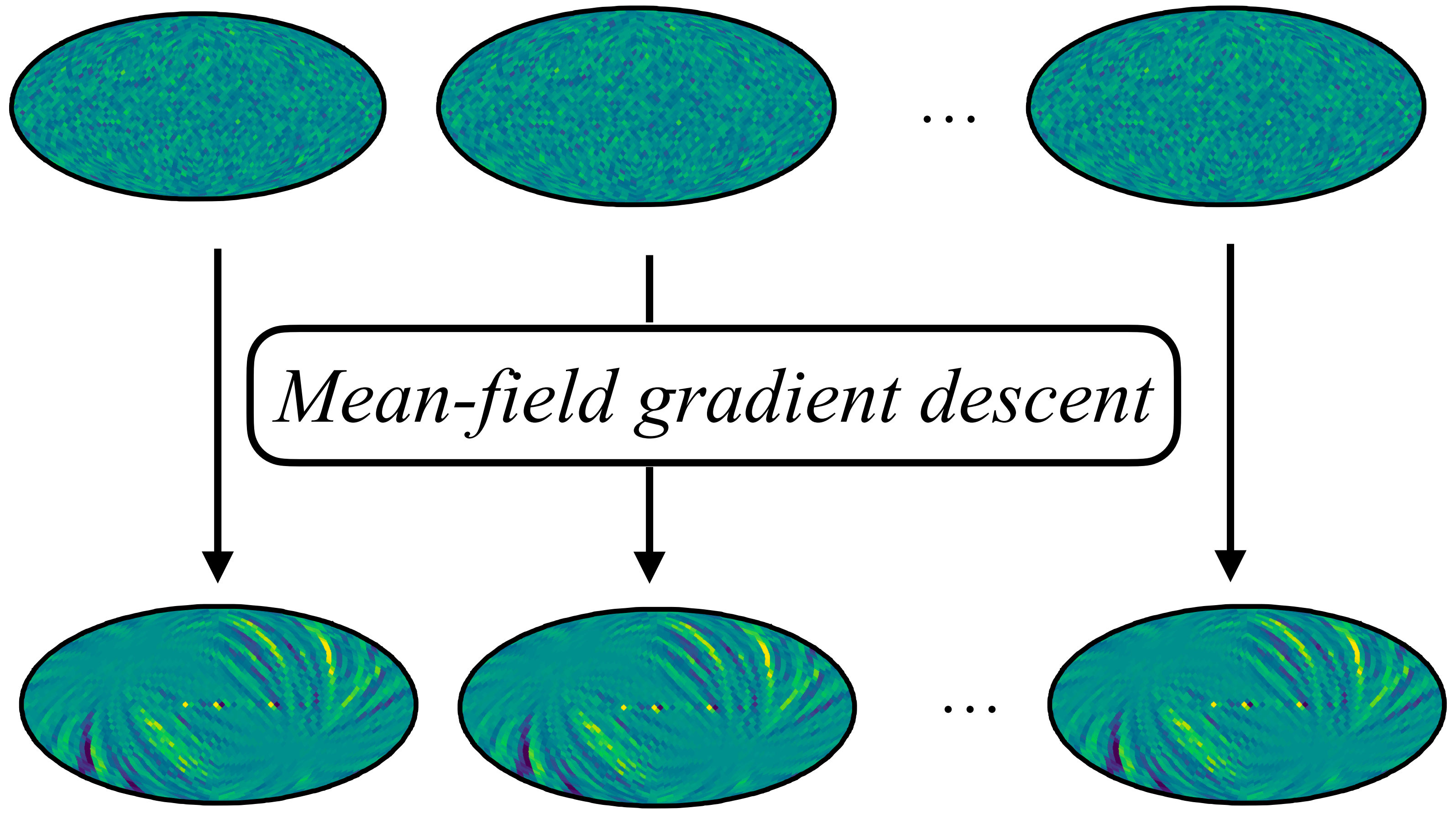
3. Emulation

A. Take batch of *white noise* samples $\{x_j\}$

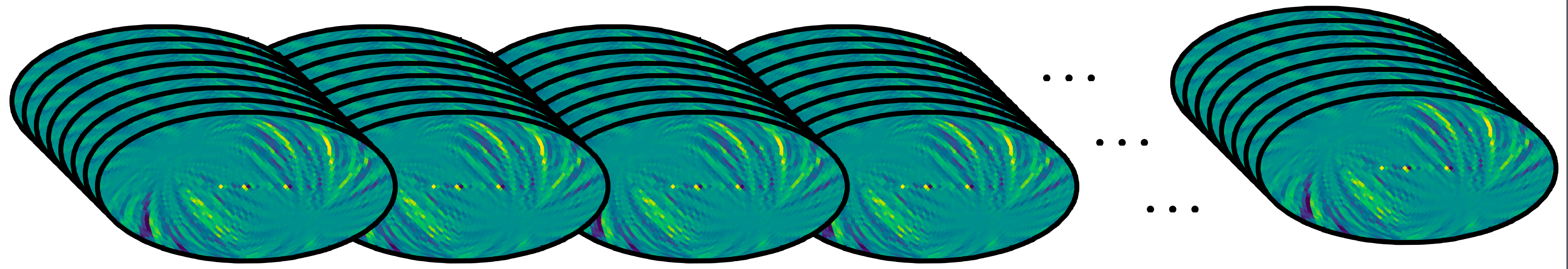
B. Minimize *loss* w.r.t. latent target in pixel space:

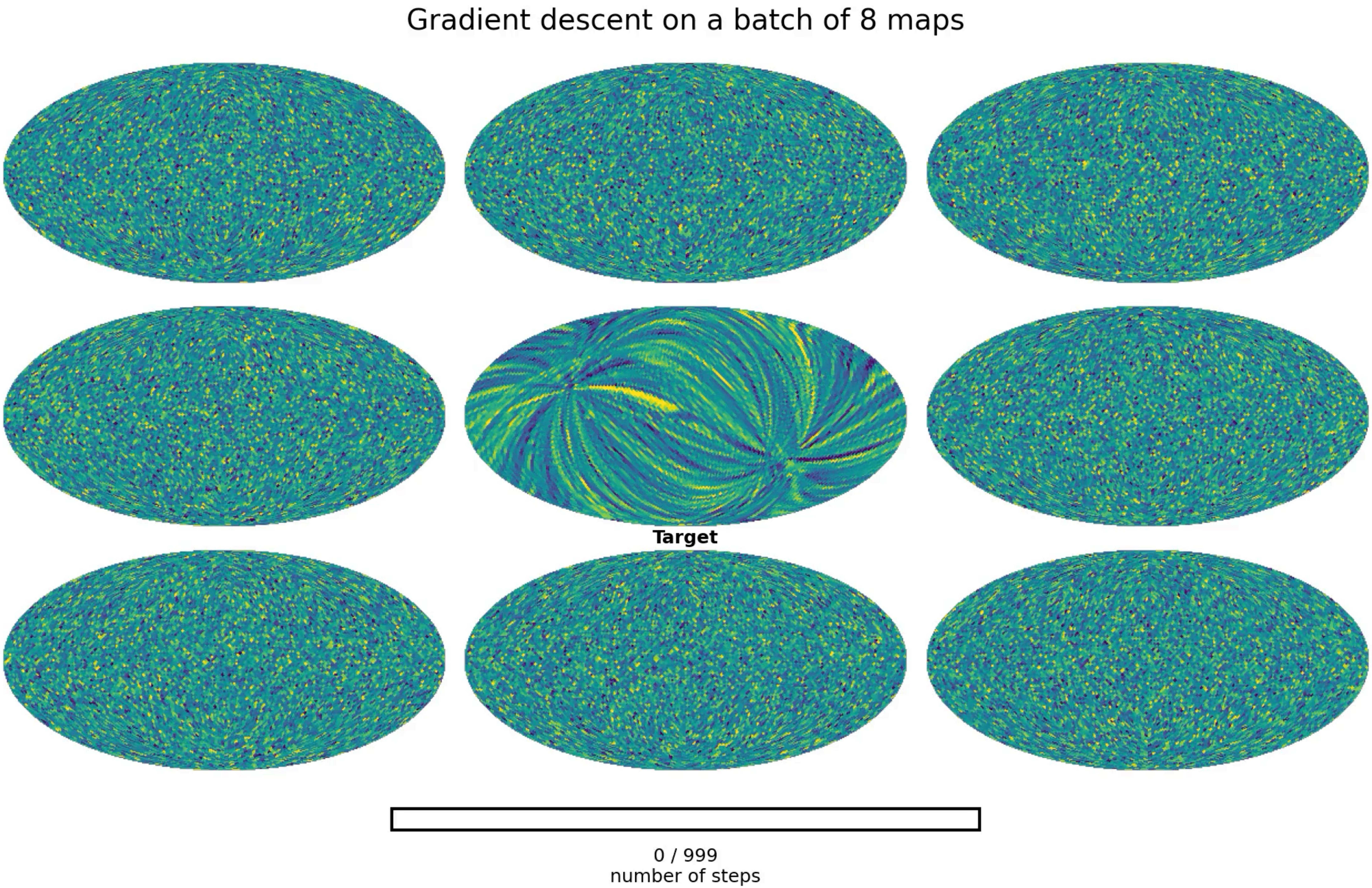
$$\mathcal{L} = \|\operatorname{Ave}_j \Phi(x_j) - z_i\|_2^2$$

C. Get batch of *emulated samples*

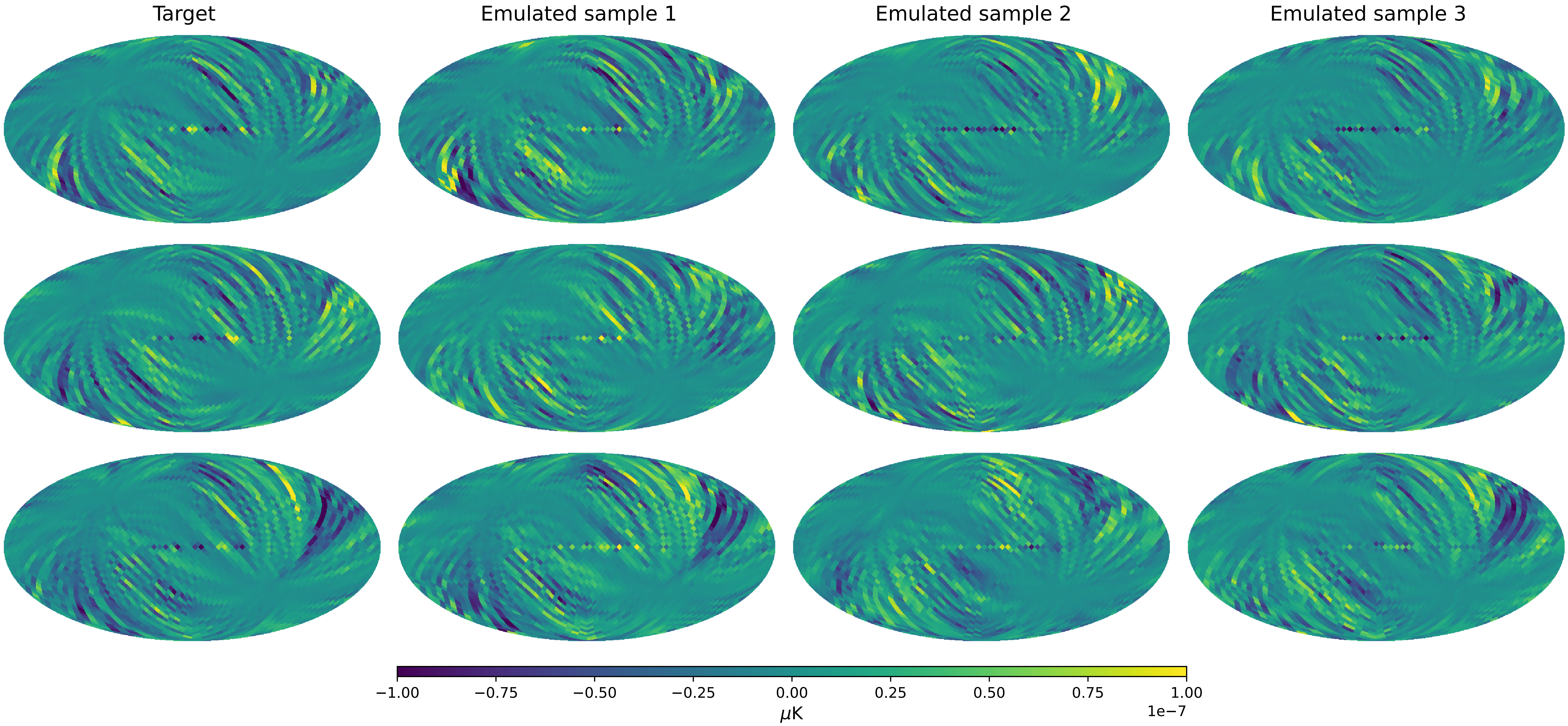


Repeat steps A-C to get many emulations




Gradient descent on a batch of 8 maps
Target
0 / 999
number of steps

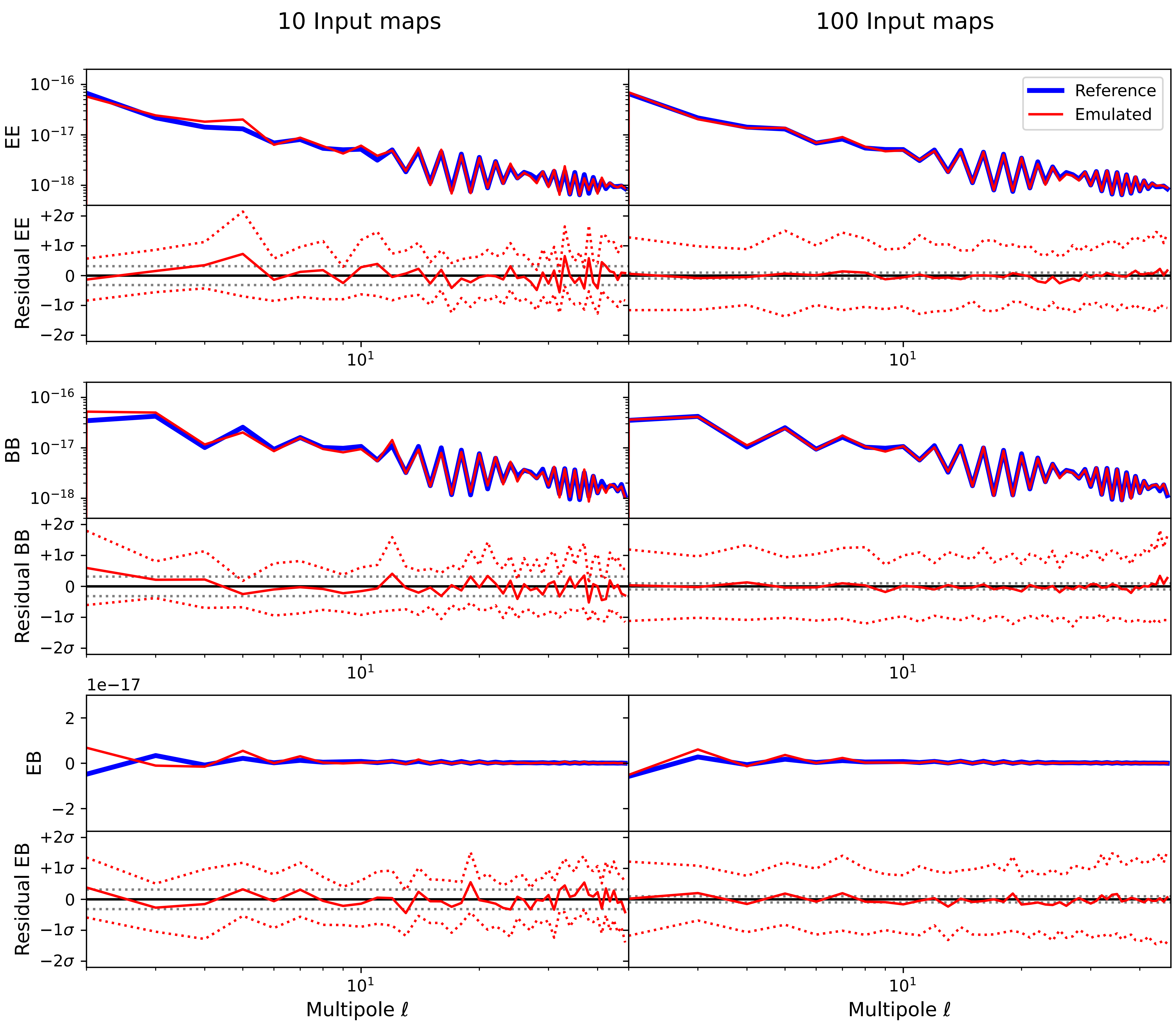
Main Results - Application to *Planck*-like scanning with additive Gaussian random *gain miscalibration systematic*



Finanziato dall'Unione europea NextGenerationEU | Ministero dell'Università e della Ricerca | Italiadomani Piano Nazionale di Ripresa e Resilienza | ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Main Results: Validation on Power Spectra

- Blue: Validation set (10,000 sims)
- Red: Emulated set (10,000 emulations)
- Gray band: std of validation set
- Residual solid red: $\dfrac{\langle C_\ell^{\mathrm{emu}}\rangle - \langle C_\ell^{\mathrm{val}}\rangle}{\sigma_\ell^{\mathrm{val}}}$
- Dotted red: $\dfrac{\langle C_\ell^{\mathrm{emu}}\rangle - \langle C_\ell^{\mathrm{val}}\rangle \pm \sigma_\ell^{\mathrm{emu}}}{\sigma_\ell^{\mathrm{val}}}$
- Dashed grey: error on the mean $\pm\dfrac{\sigma_{\mathrm{val}}}{\sqrt{N_{\mathrm{input}}}}$



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Main results - Pixel-Pixel Correlation Matrices from 100 input maps

Finanziato dall'Unione europea
NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC
Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Main results - Pixel-Pixel Correlation Matrices from 100 input maps

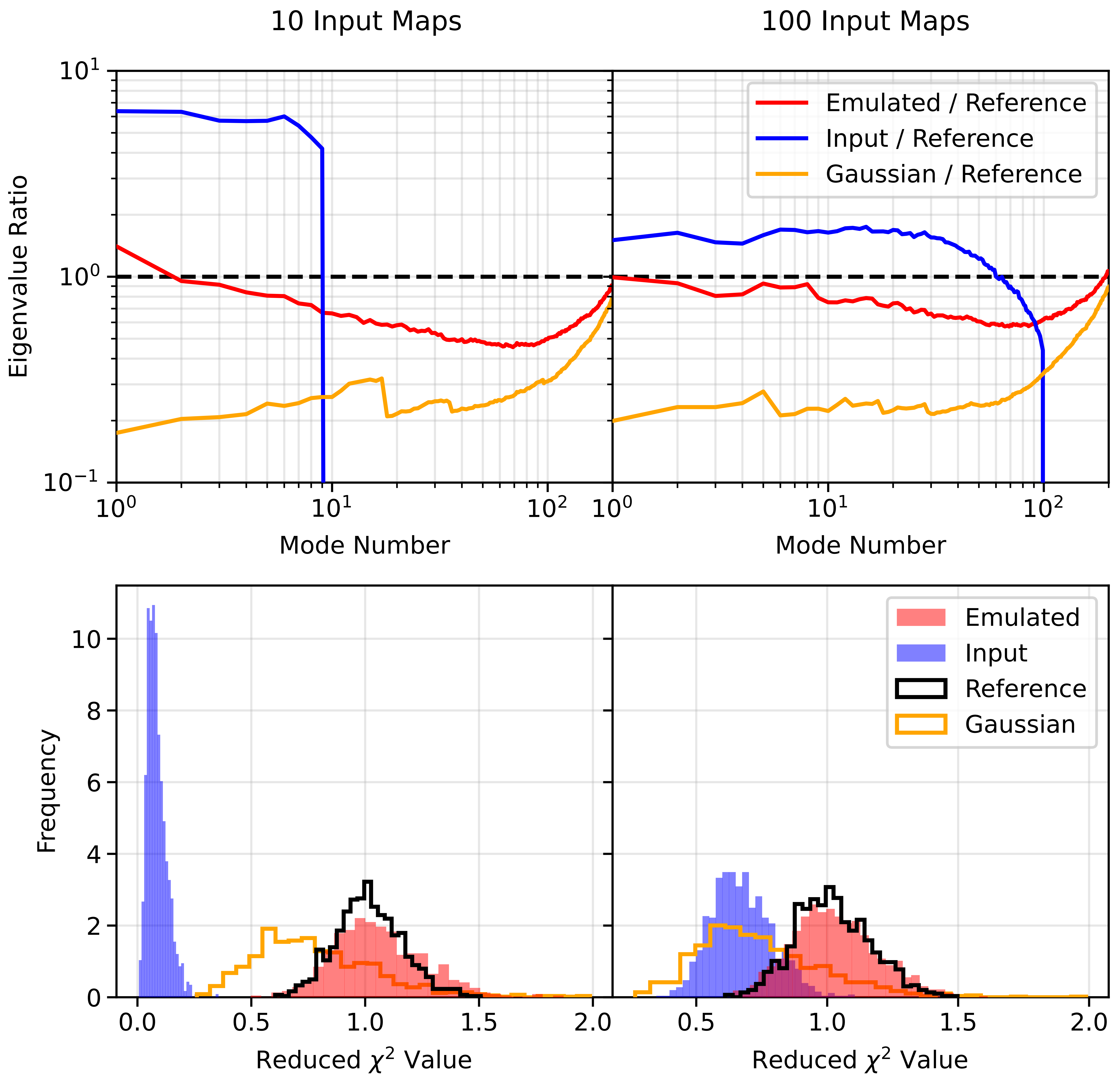
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Main results - Eigenvalues and χ^2

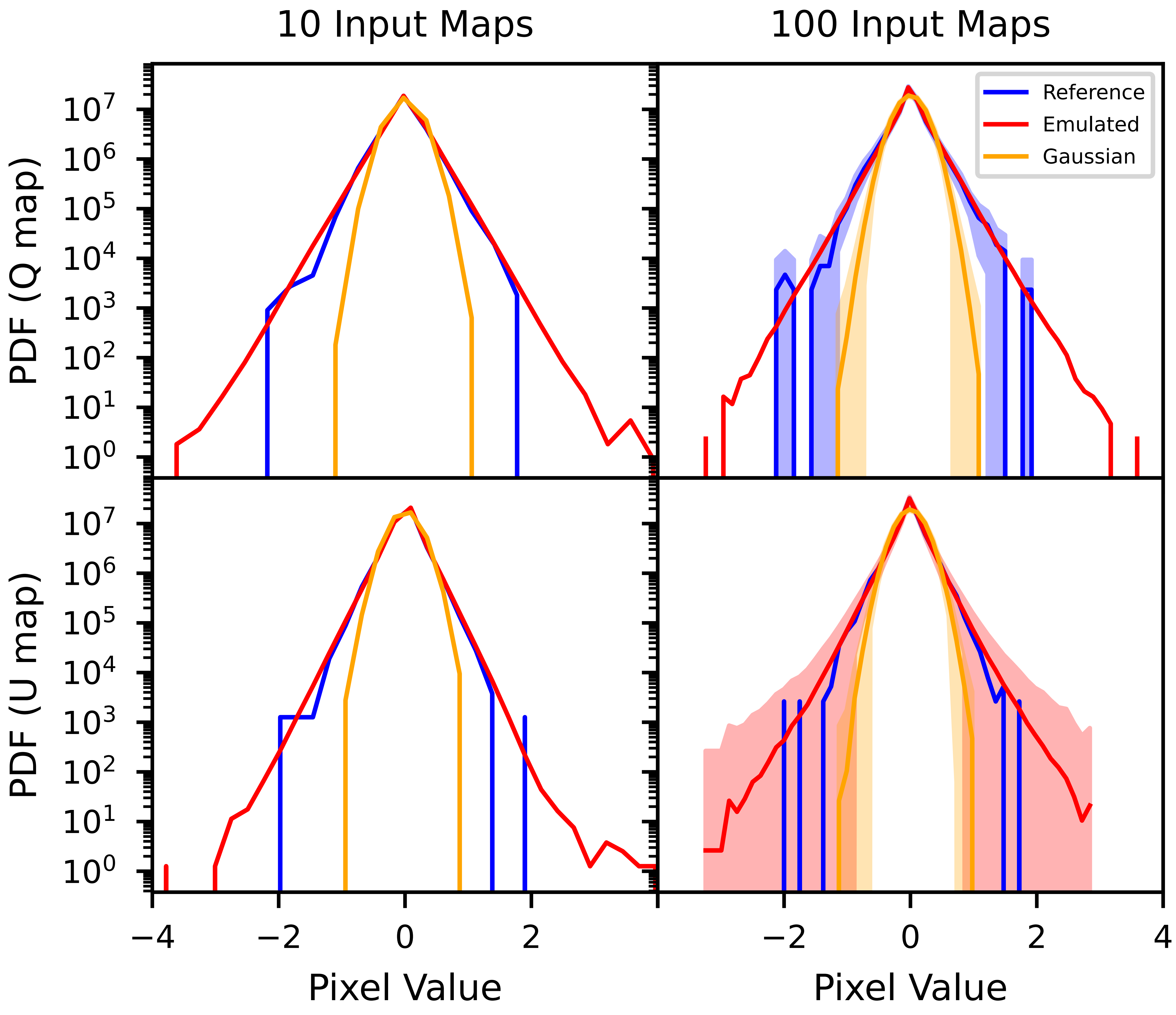
- Reduced χ^2 histogram:

$$\chi^2 = \mathbf{m}^T \mathbf{C}^{-1} \mathbf{m}/d$$

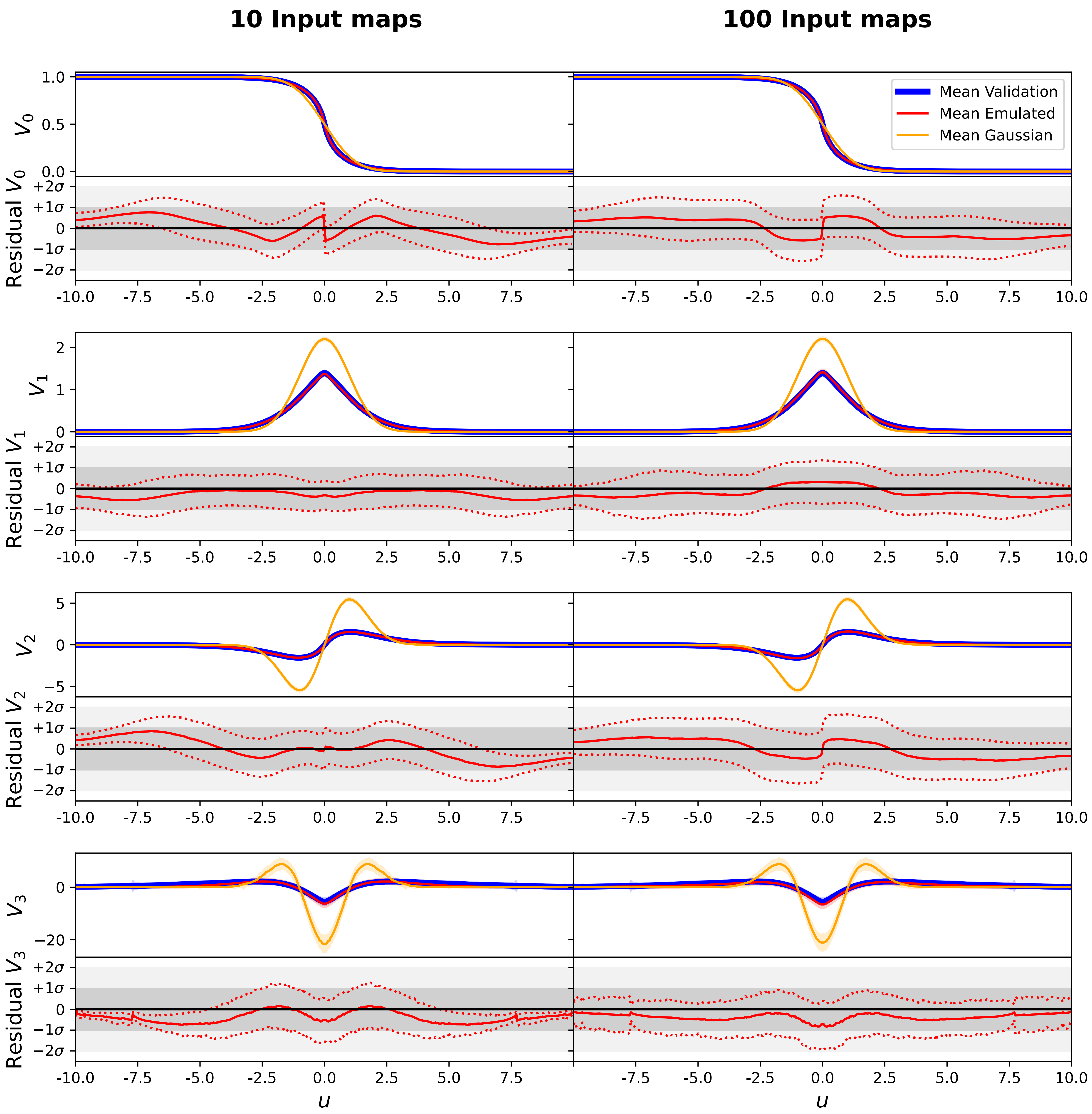
- **m**: mean subtracted validation set maps
- **C** either reference, input or emulated pixel covariance
- Compare also to naive Gaussian realizations from isotropic power spectrum



Main results - Validation on PDFs of maps

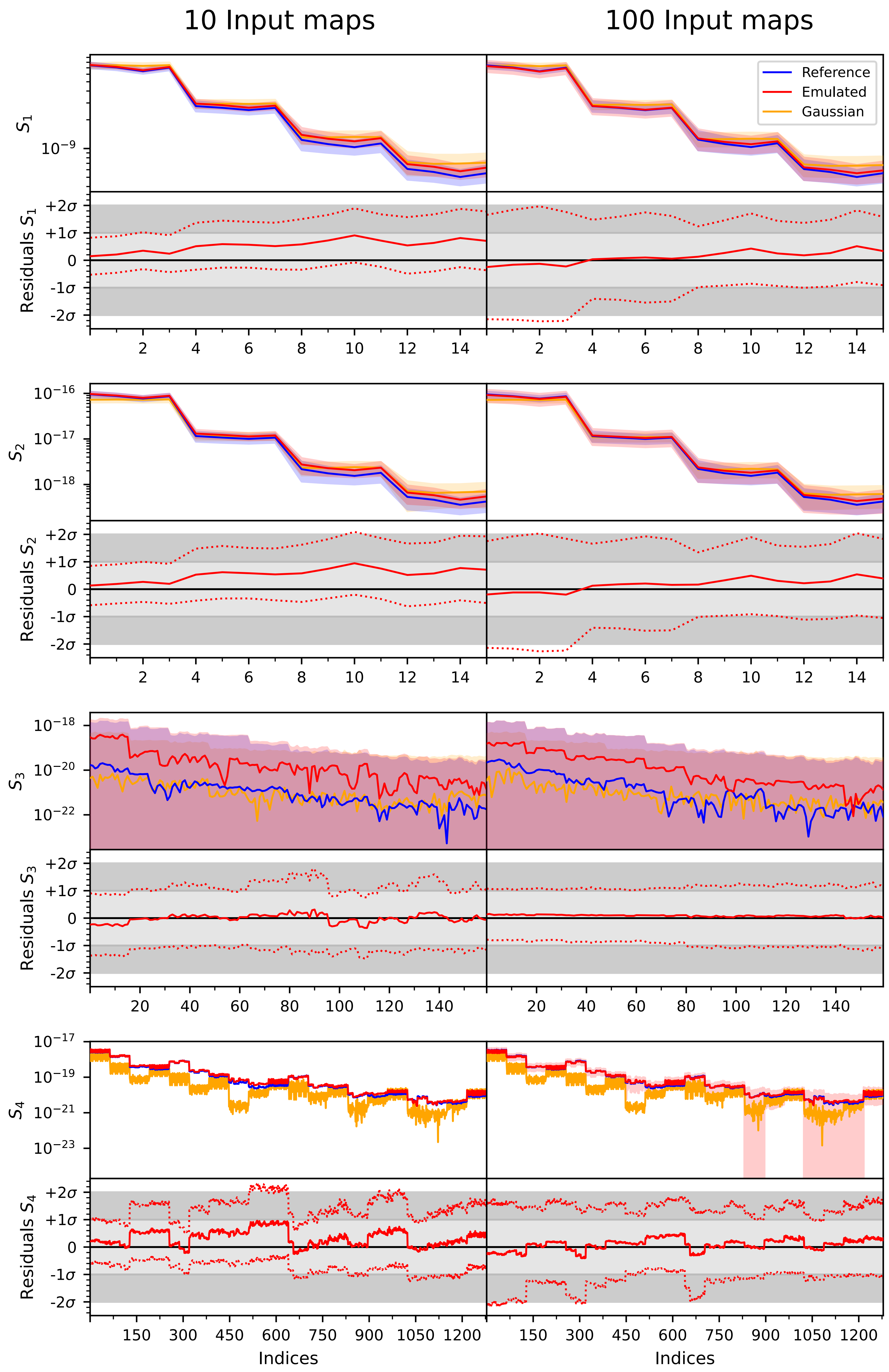


Main results - Minkowski Functionals



Main results - Scattering Coefficients

- Kernel 3x3
- 4 orientations
- 4 scales ($J_{\max} = 4$)
- Dyadic
- Total of 1474 parameters in scattering coefficients:
  - S_0 : 2
  - S_1: 4 x 4 =16
  - S_2: 4 x 4 =16
  - S_3: 10 x 4 x 4 =160
  - S_4: 20 x 4 x 4 x4 =1280



Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Next Steps and Expected Results

Computational benchmark

- Emulator is ~10^4 times faster than going through TOD simulation and runs
- Extremely efficient at low-resolution ($N_{\mathrm{side}} \lesssim 256$)
- Still very fast but too memory-hungry for high resolution ($N_{\mathrm{side}} \gtrsim 512$) → we're working on that

Software

- HEALPIXML: Python+TensorFlow/Torch software for scattering transform:
  - Runs on single/multiple GPUs
  - Available on **GitHub**
- CMBSCAT Emulator + Jupyter notebook demo available on **GitHub**
- Paper out arxiv:2503.11643 submitted to A&A, extremely positive report

Future Applications

- Applications:
  - Killing sample variance in covariance matrices
  - *Simulation-based inference*
  - *Denoising* of complex instrumental systematics
  - …
- Completion rate 85-90% exceeding expectations