NEURIPS 2026 · IMPERIAL COLLEGE LONDON

# Mamba Flow Matching Neural Processes: Linear-Time Inference for Irregularly Observed Spatial Fields

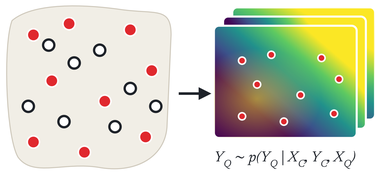
**Cosmo Santoni**, Giovanni Charles, Timothy James Hitge, Oliver John Watson, Elizaveta Semenova

Code & Paper
github.com/CosmoNaught/MambaFlowNP

## Mapping a continent from scattered stations, *in just a second.*

## 1 Neural processes

Context and query locations → Joint predictive draws




Interpolating a spatial field, such as daily maximum temperature, from **sparse and irregularly located observations** is a recurring task across the environmental and health sciences.

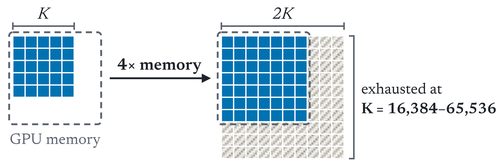
The standard tool, **Gaussian-process kriging**, refits hyperparameters and solves an **O(N³) system for every new field**. **Neural processes (NPs)** instead learn the interpolation map from past fields and predict a new field in a **single forward pass**.

We propose **MambaFlowNP**, a joint-predictive NP with **linear compute**, trained with a **conditional flow-matching** objective.

KEY
- **context point** ($X_C$, $Y_C$)
- **query location** $X_Q$
- **hidden state** at a token position

## 2 Joint-predictive NPs

**Self-attention**



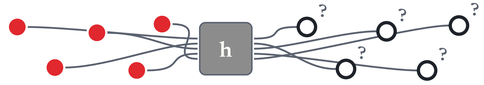

**Mamba-2 scan**




Joint-predictive NPs such as TNPA, NDP and FNP use **self-attention** over all K = N + M context and query points, with compute scaling as **O(K²)**.

### One hidden state for all context

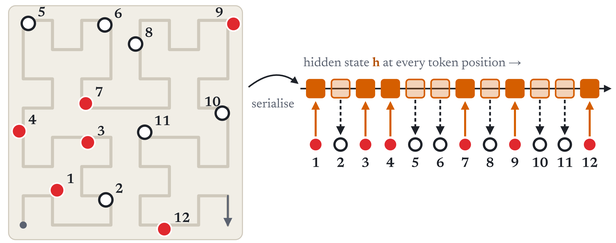
**Naive Mamba-2**




State-space models such as **Mamba-2** run a recurrent scan in **O(K)** time, FLOPs and activation memory. **However**, a naive adaptation compresses the context into a **single hidden state**, a bottleneck that raises MSE by **58–65% on NOAA**.

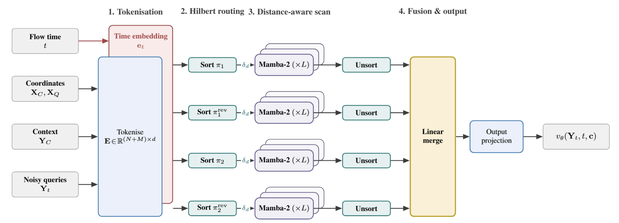
## 3 MambaFlowNP (MFNP)

**a** Hilbert serialisation → **b** One shared sequence, Mamba-2 scan



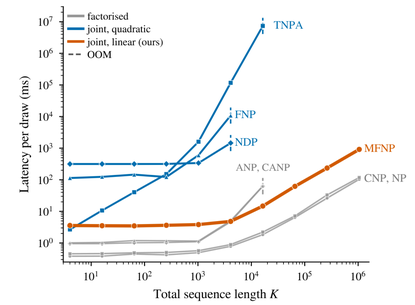

Hilbert curves order the combined context and query points so that nearby locations tend to remain close in the sequence; queries enter the scan with zero input and zero step size.

### Architecture




1. **Tokenise.** Frozen Random Fourier Feature (RFF) coordinate embeddings
2. **Route.** Two Hilbert orderings (original and swapped axes) and their reversals
3. **Scan.** Distance-aware Mamba-2 scans modulated by an adaLN-Zero flow-time embedding $e_t$
4. **Fuse and read out.** Inverse-permutation fusion to per-query velocities

## 4 Computational scaling

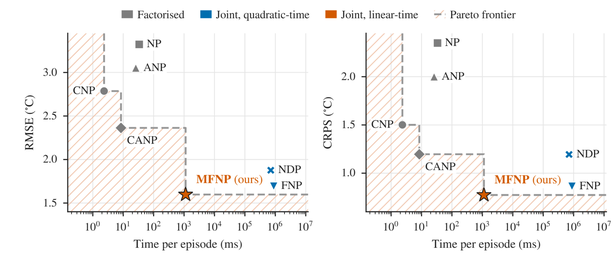



**$10^6$ points**
one joint predictive draw in **913.86 ms** on one GPU

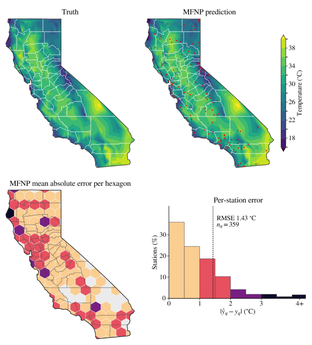
**114–784×**
**lower latency** than the strongest joint-predictive baseline on NOAA

**0.955**
**Cov-95** on continental US NOAA, nominal 0.95

## 5 Pareto frontier at the *continent scale*




### California 122 context stations · 364 predicted

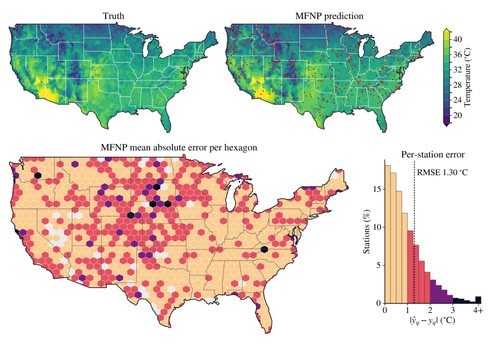



**1.344** °C
**lowest RMSE**
Next best: TNPA 1.378

**0.945**
**Cov-95**, nominal 0.95
Next best: TNPA 0.938

**67.9** ms
**per episode**
Next best: FNP 8,844 ms

### Continental US 558 context stations · 5,022 predicted




**1.598** °C
**lowest RMSE**
Next best: FNP 1.695

**0.955**
**Cov-95**, nominal 0.95
Next best: FNP 0.786

**1.13** s
**per episode**
Next best: NDP 708 s

### NOAA, joint-predictive NPs

| Model | RMSE (°C) ↓ | CRPS (°C) ↓ | Cov-95 | Time (ms) ↓ |
|---|---|---|---|---|
| *California (K = 486)* | | | | |
| TNPA | 1.378 | 0.701 | 0.938 | $3.004\times10^4$ |
| NDP | 1.753 | 1.064 | 0.590 | $9.154\times10^3$ |
| FNP | 1.548 | 0.839 | 0.785 | $8.844\times10^3$ |
| **MFNP** | **1.344** | **0.681** | **0.945** | **67.9** |
| *Continental US (K = 5,580)* | | | | |
| TNPA | | *T/O* | | |
| NDP | 1.878 | 1.194 | 0.322 | $7.075\times10^5$ |
| FNP | 1.695 | 0.864 | 0.786 | $8.831\times10^5$ |
| **MFNP** | **1.598** | **0.774** | **0.955** | $\mathbf{1.126\times10^3}$ |

Mean over 10 evaluation seeds; time per episode ($N_{samples}$ = 128) on one GPU.

### Limitations

- MFNP targets irregular point sets; on regularly gridded domains, convolutional NPs and FFT-based GP solvers are typically more efficient.
- At small K ($\lesssim 10^3$), the per-call dispatch floor leaves MFNP slower than the simplest factorised NPs.
- One- and two-dimensional inputs only; space–time inputs are left to future work.
- Real-data evaluation is limited to NOAA daily maximum temperature, with a temporal split on a fixed station network.
- Predictive quality at $10^6$ points is unestablished; speedups combine reflow distillation and the linear-time scan.

Funded by Wellcome, MRC GIDA and Schmidt Sciences (AI2050).