# Fourier neural operator algorithm details

***Building a machine learning climate emulator***
**Harvard Earth and Planetary Science 226, Eli Tziperman**

The model first appends latitude and longitude coordinates to the six input maps and uses a two-layer fully connected network to "lift" the eight values at each grid cell into 32 learned features. It then pads the Pacific domain to make it consistent with the periodic boundaries assumed by the Fourier transform. Four Fourier blocks combine a global spectral path, which mixes retained large-scale Fourier modes across all channels, with a local pointwise path. After the blocks, the padding is removed and a second pointwise two-layer fully connected network projects the 32 features back to three output fields. The result is next-month maps of SST, D20, and zonal wind stress.

## 1. State and model input

### 1.1 State, normalization, and calendar

Let $c \in \{1,2,3\}$ denote SST, D20, and zonal wind stress; let $M_{ij} \in \{0,1\}$ be the fixed Pacific mask on the $41 \times 170$ grid, $\Omega = \{(i,j) : M_{ij} = 1\}$, and $N = 5810$. From the first $8400 = 700 \times 12$ months, one scalar mean and population standard deviation are stored per field,

$$\mu_c = \frac{1}{8400N} \sum_{t=0}^{8399} \sum_{(i,j)\in\Omega} F_{tcij}, \qquad \sigma_c = \left[ \frac{1}{8400N} \sum_{t=0}^{8399} \sum_{(i,j)\in\Omega} (F_{tcij} - \mu_c)^2 \right]^{1/2}. \tag{1}$$

At month $t$, define $S_{tcij} = (F_{tcij} - \mu_c)/\sigma_c$ on $\Omega$ and $S_{tcij} = 0$ outside $\Omega$. Let $m_t \in \{0,\dots,11\}$ be the zero-based calendar month. Tensor dimensions below omit the batch axis. During deterministic inference, batch members are evaluated independently using the same learned parameters.

### 1.2 Coordinate-augmented grid input

The six supplied channels are

$$X_t[:,i,j] = \left(S_{t1ij}, S_{t2ij}, S_{t3ij}, \sin(2\pi m_t/12), \cos(2\pi m_t/12), M_{ij}\right)^\top \in \mathbb{R}^6. \tag{2}$$

Fixed grid coordinates $r_i = i/40$ and $s_j = j/169$ are appended to the six input channels. Thus $z_{ij} = [X_t[:,i,j]; r_i; s_j] \in \mathbb{R}^8$.

## 2. Network building blocks

### 2.1 Pointwise lifting and symmetric domain padding

The activation function GELU operates elementwise:

$$\mathrm{GELU}(x) = x\Phi(x) = \frac{x}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right]. \tag{3}$$

There is no batch/layer normalization and no dropout. The two-layer fully connected network for channel lifting is shared at all grid cells and lifts eight values to 32 features through 64 hidden values:

$$a_{ij} = \mathrm{GELU}(Az_{ij} + a), \qquad h^{(0)}_{ij} = Ba_{ij} + b, \quad A \in \mathbb{R}^{64\times 8}, \quad B \in \mathbb{R}^{32\times 64}. \tag{4}$$

Here $h^{(0)}$ is the lifted input grid: each grid cell is represented by 32 learned features, and the superscript (0) indicates that it is the input to the first Fourier block. After lifting, the model adds zero padding equal to $\sim 10\%$ of each spatial dimension on each side: four zero rows on both latitude sides and 17 zero columns on both longitude sides, so $h^{(0)} \in \mathbb{R}^{32\times 49\times 204}$. This padding is removed from the model output after processing by all four Fourier blocks.

### 2.2 Spectral convolution path

Each Fourier block has two parallel paths: the spectral convolution path uses low-frequency Fourier coefficients to represent global relationships across the padded grid, while the local $1 \times 1$ convolution bypass applies a learned linear transformation independently at each grid cell. The two path outputs and the spectral bias are added before GELU is applied in the first three blocks. For block $\ell$, let $u = h^{(\ell-1)}$ and let $\widehat{u} = \mathcal{F}u$ be the real 2-D FFT with forward normalization. Thus, for each channel $d$,

$$\widehat{u}_{dkq} = \frac{1}{49 \cdot 204} \sum_{i=0}^{48} \sum_{j=0}^{203} u_{dij} \exp\left[-2\pi \mathrm{i}\left(\frac{ki}{49} + \frac{qj}{204}\right)\right]. \tag{5}$$

Along latitude, both negative and positive frequencies are retained and reordered around zero. Along longitude, only zero and positive frequencies are stored because the omitted negative-frequency coefficients can be reconstructed using the symmetry of the two-dimensional FFT. The model truncates the Fourier representation, retaining only 16 low-frequency latitude modes $k = -8,\dots,7$ and 17 nonnegative longitude modes $q = 0,\dots,16$.

This assumes that these large-scale modes are sufficient to represent the dynamics of interest. For each retained pair $(k, q)$ and output channel $p$, $\widehat{v}^{(\ell)}_{pkq}$ is the output Fourier coefficient produced by Fourier block $\ell$. $R^{(\ell)}_{dpkq}$ is a learned complex-valued weight in Fourier block $\ell$.

$$\widehat{v}^{(\ell)}_{pkq} = \sum_{d=1}^{32} R^{(\ell)}_{dpkq} \widehat{u}_{dkq}, \qquad R^{(\ell)} \in \mathbb{C}^{32\times32\times16\times17}; \quad \widehat{v}^{(\ell)}_{pkq} = 0 \text{ outside the retained set.} \tag{6}$$

The inverse transform returns a real grid and a learned constant spectral bias $c^{(\ell)}_p$ is added.

### 2.3 Local bypass path

In parallel, a bias-free $1 \times 1$ convolution applies a learned linear transformation to the 32 features independently at each grid cell:

$$\left(P^{(\ell)}u\right)_{pij} = \sum_{d=1}^{32} P^{(\ell)}_{pd} u_{dij}, \qquad P^{(\ell)} \in \mathbb{R}^{32\times32}. \tag{7}$$

### 2.4 Combining the two paths

With $\mathcal{F}^{-1}$ denoting the inverse Fourier transform, the two paths are combined and the output of block $\ell$ is computed as

$$g^{(\ell)} = \mathcal{F}^{-1}\left(R^{(\ell)}\,\mathcal{F}h^{(\ell-1)}\right) + c^{(\ell)} + P^{(\ell)}h^{(\ell-1)}, \qquad h^{(\ell)} = \begin{cases} \mathrm{GELU}(g^{(\ell)}), & \ell = 1, 2, 3, \\ g^{(4)}, & \ell = 4. \end{cases} \tag{8}$$

Every retained output Fourier coefficient is computed from all 32 input channels, and every output cell depends on the entire padded grid. Before the inverse real FFT, the omitted negative-frequency coefficients are reconstructed as complex conjugates of the corresponding positive-frequency coefficients, ensuring a real-valued output grid.

## 3. Layer-by-layer evaluation

The padded spectral features are returned to the physical grid and projected to the three predicted fields as follows.

### 3.1 Network evaluation and output

After block 4, the padding (four rows and 17 columns on each side) is removed, restoring $h^{(4)} \in \mathbb{R}^{32\times41\times170}$. Finally, a second two-layer fully connected network is applied independently at every grid cell, converting the 32 features at that cell into the three output fields:

$$d_{ij} = \mathrm{GELU}(Ch^{(4)}_{ij} + c), \qquad Y_{t,:,i,j} = Dd_{ij} + d, \qquad C \in \mathbb{R}^{64\times32}, \quad D \in \mathbb{R}^{3\times64}. \tag{9}$$

Although only low Fourier modes are learned explicitly, the GELUs between blocks can transfer information among resolved modes, while the local bypass preserves grid-scale structure.

### 3.2 Dimensions and parameter accounting

| stage | input | exact operation | output | parameter elements |
|---|---|---|---|---|
| grid embedding | $6 \times 41 \times 170$ | append $r_i, s_j$ | $8 \times 41 \times 170$ | 0 |
| lifting | $8 \times 41 \times 170$ | pointwise $8 \to 64 \to 32$ | $32 \times 41 \times 170$ | 2,656 |
| padding | $32 \times 41 \times 170$ | symmetric zeros: 4, 17 per side | $32 \times 49 \times 204$ | 0 |
| each of 4 blocks | $32 \times 49 \times 204$ | spectral + local $1 \times 1$ | same | 279,584 |
| unpadding | $32 \times 49 \times 204$ | retain physical grid | $32 \times 41 \times 170$ | 0 |
| projection | $32 \times 41 \times 170$ | pointwise $32 \to 64 \to 3$ | $3 \times 41 \times 170$ | 2,307 |
| | | total trainable parameter elements | | **1,123,299** |

Each block contains 278,528 learned complex spectral coefficients, 32 real spectral biases, and 1,024 real local weights. Counting one complex coefficient as one parameter element, the full model has 1,114,112 complex and 9,187 real elements, equivalent to 2,237,411 real scalars.

## 4. Deterministic monthly recurrence

Starting from a standardized state $S_t$, construct the model input described in Section 1 and evaluate the network to obtain $Y_t = f_\theta(S_t, m_t; M)$, with no output activation. During a deterministic free run, impose the ocean mask and advance the calendar:

$$S_{t+1,cij} = M_{ij}Y_{t,cij}, \qquad m_{t+1} = (m_t + 1) \bmod 12. \tag{10}$$

Rebuild the model input from the new state and calendar month and repeat the complete network. Saved ocean output is converted to physical units field by field:

$$\widehat{F}_{t+1,cij} = \sigma_c S_{t+1,cij} + \mu_c \quad ((i, j) \in \Omega). \tag{11}$$

Land is written as missing (NaN). The only state carried between forecast months is the three-channel standardized prediction; no hidden features are carried forward. The calendar advances analytically, while the mask and grid remain fixed.

## 5. Architecture at a glance

