Practical Introduction to Wavelets

7th Summer School on Theory, Mechanisms and Hierarchical Modelling of Climate Dynamics.

Estimating Ocean Transports: Single Sections, Box Models and Reanalysis Products

Paulo S. Polito

Instituto Oceanográfico — Universidade de São Paulo

The Abdus Salam International Centre for Theoretical Physics

IAEA International Atomic Energy Agency

unesco United Nations Educational, Scientific and Cultural Organization

Objectives

- Place the RAPID-MOCHA mooring array in its oceanographic context.
- Explore raw hydrographic time series and diagnose their limitations.
- Remove the deterministic background (mean, linear trend, and seasonal cycle) to isolate anomalies.
- Apply the Fast Fourier Transform (FFT) to quantify how variance is distributed across timescales.
- Understand the fundamental trade-off between time and frequency that motivates wavelet analysis.

Prerequisites

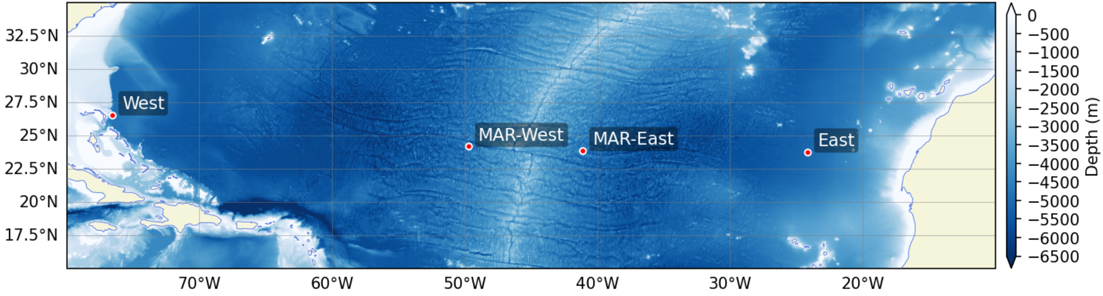
- Basic Python: `NumPy, masked arrays, dictionaries, Matplotlib`.
- Elementary statistics: mean, variance, linear regression.
- Trigonometric identities and basic calculus.
- Familiarity with NetCDF files as the standard data format - at least for us.
- An environment with basic Python plus `netCDF4, gsw, Cartopy`, and `pycwt`.

Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. Numerical Intuition
6. The Continuous Wavelet Transform (CWT)

The RAPID-MOCHA Observing System: `Rapid_1_map.py`

- The RAPID-MOCHA array monitors the Atlantic Meridional Overturning Circulation (AMOC) continuously since 2004 at $\approx 26.5°$N.
- Instruments sample top to bottom T and S at 12-hour sampling rate.
- West, MAR-West/East and East sites: shelf-break, Ridge and abyssal plain.



Geospatial Context

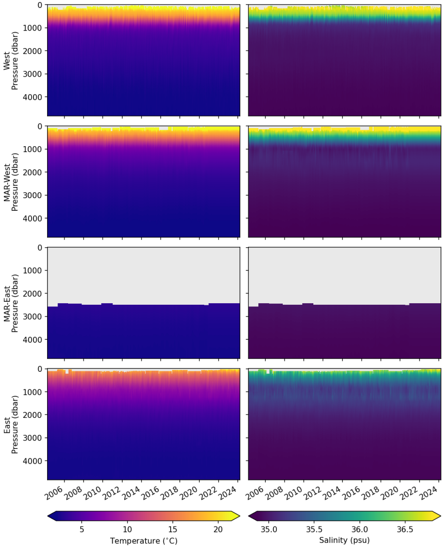
- Dataset: ETOPO 2022 global relief model at 1 arc-minute resolution (`NETCDF4` format).
- Big data strategy: use `np.searchsorted` (binary search, $\mathcal{O}(\log N)$) to read *only the geographic slice* needed—never load the full global grid.
- True Map rendered with Cartopy, boring Plate Carrée projection.
- Look at the bathymetric gradient: the West site sits right on a steep slope. What dynamics does that imply?

Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. Numerical Intuition
6. The Continuous Wavelet Transform (CWT)

Raw Temperature and Salinity Profiles: `Rapid_2_read.py`

- 4 sites, $T(t,p)$, $S(t,p)$ sections.
- Shared axes (`sharex, sharey`): zoom on one panel updates all seven others.
- Color limits set statistically. Why?
- Gray background $\rightarrow$ NaN as missing data.
- What are the three problems immediately visible?



Three Problems to Solve Before Any Analysis

1. Surface gap: moorings do not reach the sea surface (shipping traffic, wave action). Data is missing in the top $\sim 30\,\text{dbar}$.

   Fix: extend the shallowest valid measurement upward to $0\,\text{dbar}$.

2. Overwhelming seasonality: the upper ocean thermohaline cycle repeats every year, swamping the interannual AMOC signal.

   Fix: fit and subtract the mean + trend + annual harmonic simultaneously.

3. MAR-East reliability: large gaps throughout the record.

   Fix: discard MAR-East entirely. Bad data is worse than no data.

Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. Numerical Intuition
6. The Continuous Wavelet Transform (CWT)

Gap Filling Strategy: Rapid_3_anomaly_calc.py

- Simple but robust: replicate the shallowest valid sensor upward.
- `np.argmax(~mask, axis=0)` finds the 1st valid depth at every timestep.
- Broadcasting: a 2-D index grid (`p_idx = np.arange(n_p)[:, np.newaxis]`) avoids slow t loop.
- The surface values are *set*, not interpolated.

```
def fill_top_gaps(arr):
    filled      = arr.copy()
    filled.mask = np.ma.getmaskarray(filled)
    n_p, n_t    = filled.shape
    % index of first valid depth (t-axis)
    first_valid = np.argmax(
        ~filled.mask, axis=0)
    p_idx  = np.arange(n_p)[:,np.newaxis]
    to_fill = p_idx < first_valid
    if not to_fill.any():
        return filled
    shallow_val = filled.data[
        first_valid, np.arange(n_t)]
    filled.data[to_fill] = \
        shallow_val[np.where(to_fill)[1]]
    filled.mask[to_fill] = False
    return filled

```

Least-Squares Decomposition

- At every pressure level i we fit a four-parameter model:

$$z^*(t) = a + b\,t + c\sin\left(\frac{2\pi t}{P}\right) + d\cos\left(\frac{2\pi t}{P}\right), \quad P = 365.25 \text{ days}$$

- The design matrix $\mathbf{A}$ (dimension $N_t \times 4$) encodes all 4 basis functions:

$$\mathbf{A} = \begin{bmatrix} 1 & t_1 & \sin(\omega t_1) & \cos(\omega t_1) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_N & \sin(\omega t_N) & \cos(\omega t_N) \end{bmatrix}$$

- Solve for $[a, b, c, d]$ simultaneously, avoiding the "one extra summer" bias:

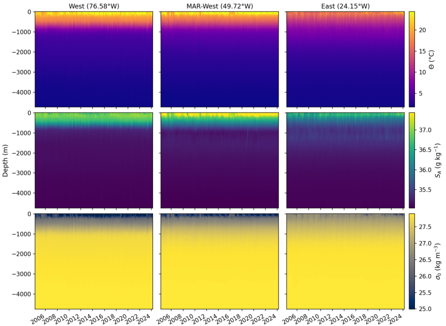
$$z^* = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top z$$

GSW/TEOS-10 Thermodynamic Conversions

- Raw mooring measurements: in-situ temperature T and practical salinity S.
- Python library `gsw`, the Gibbs SeaWater Toolbox, a.k.a. TEOS-10, converts these to thermodynamically consistent variables:
  - S_A — Absolute Salinity ($\mathrm{g\,kg^{-1}}$), accounts for regional composition anomalies.
  - Θ — Conservative Temperature (°C), proportional to ocean heat content.
  - σ_0 — Potential density anomaly ($\mathrm{kg\,m^{-3}}$), referenced to 0 dbar.
- From this point on, σ_0 is our primary dynamical variable: it directly governs buoyancy and baroclinicity.

Original Fields (Θ, S_A, σ_0): `Rapid_4_anomaly_plot.py`

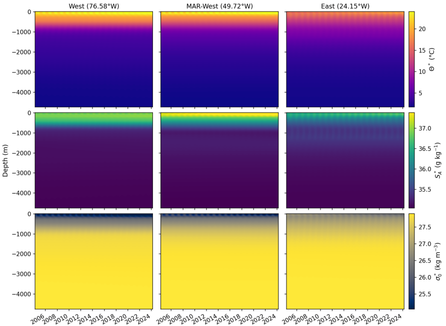
- Sequential colormaps for total fields: 0 isn't special.
- The vertical gradient dominates, with the thermocline dipping westward. Why?



- MAR-West shows yellow regions: sensor failures and the surface gap filled upwards.
- Can you identify the seasonal signal superimposed on the stratification?

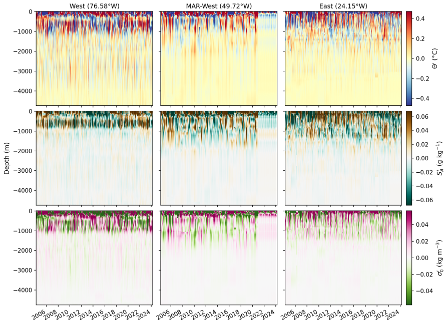
Mean + Trend + Seasonal (Θ^*, S_A^*, σ_0^*)

- Background: vertical stratification, small surface seasonal cycle and invisible trend.
- East site shows stronger seasonal amplitude than the West site.



- The West thermocline is more pronounced and better organized — why?
- Deep water below 1000 m is nearly uniform in space and time: NADW!

Anomaly Fields (Θ', S'_A, σ'_0)



- In Divergent colormaps 0 is special.

  Both ends are symmetric, color $\Rightarrow$ sign.

- Flat regions in MAR-West:

  climatological fill, tiny anomalies from the mean.

- Horizontal bands mark the MicroCAT instrument positions. What are your eyes perceiving?

- Energy is *distributed in time*—there are clearly active and quiet periods.

Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. Numerical Intuition
6. The Continuous Wavelet Transform (CWT)

Signal Decomposition: The Fourier Premise

- Measurements are seen as: $y(t) = \overline{y(t)} + \sum_p \left[A_p \cos(\omega_p t) + B_p \sin(\omega_p t)\right]$
- Prerequisites for Fourier analysis: continuous series, smooth variability, and—crucially—stationarity.
- A signal is ergodic if any long, contiguous segment has the same statistical properties as any other.
- Goals: (1) separate the periodic part from noise; (2) distribute the variance among components.

I.e.: we want to find A_p, B_p, and $\omega_p = \dfrac{2\pi p}{T}$ for $p = 1, 2, 3, \ldots$

The Continuous Fourier Transform

- Multiply $y(t)$ by $\cos(\omega_p t)$ and $\int_0^T$, orthogonality kills all other terms:

$$A_p = \frac{2}{T}\int_0^T y(t)\cos(\omega_p t)\,dt, \qquad p = 0, 1, 2, \ldots$$

$$B_p = \frac{2}{T}\int_0^T y(t)\sin(\omega_p t)\,dt, \qquad p = 1, 2, 3, \ldots$$

- Equivalently in amplitude–phase form (prove that to yourself at least once):

$$C_p = \sqrt{A_p^2 + B_p^2}, \qquad \phi_p = \arctan\left(\frac{B_p}{A_p}\right)$$

- C_p is the amplitude of mode p; ϕ_p is its phase (e.g., time of maximum).

The Discrete Fourier Transform (DFT)

- For discrete data $t_n = n\,\Delta t$, $T = N\,\Delta t$:

$$y(t_n) = \frac{1}{2}A_0 + \sum_{p=1}^{N/2}\left[A_p \cos\!\left(\frac{2\pi p}{T}t_n\right) + B_p \sin\!\left(\frac{2\pi p}{T}t_n\right)\right]$$

- Coefficients are computed by discrete summation over all times n:

$$A_p = \frac{2}{N}\sum_{n=1}^{N} y_n \cos\!\left(\frac{2\pi pn}{N}\right), \qquad B_p = \frac{2}{N}\sum_{n=1}^{N} y_n \sin\!\left(\frac{2\pi pn}{N}\right)$$

- The reconstruction sums over all frequencies p.

- Nyquist frequency: $p = N/2 \Rightarrow f_{\max} = \dfrac{1}{2\,\Delta t}$ (highest resolvable frequency; $\Delta t = 0.5$ day $\Rightarrow f_{\mathrm{Nyq}} = 1$ cycle/day).

Power Spectral Density

- The Power Spectral Density (PSD) divides variances among frequencies:
  $\mathrm{PSD}(f_p) = \dfrac{2\,\Delta t}{N}\,|X_p|^2, \qquad X_p =$ DFT coefficient.
- Factor of 2: single-sided spectrum (we fold the negative frequencies onto the positive side for real-valued signals, using `scipy.fft.rfft`).
- $\Delta t/N$ normalises for sampling rate and record length $\Rightarrow$ units: (variable)$^2 \cdot$ day.
- Plotted on log-log axes: power-law behaviour appears as straight lines.
- Parseval's Theorem: $\sum_p \mathrm{PSD}(f_p)\,\Delta f = \sigma_y^2$.

FFT Setup: `Rapid_5_FFT.py`

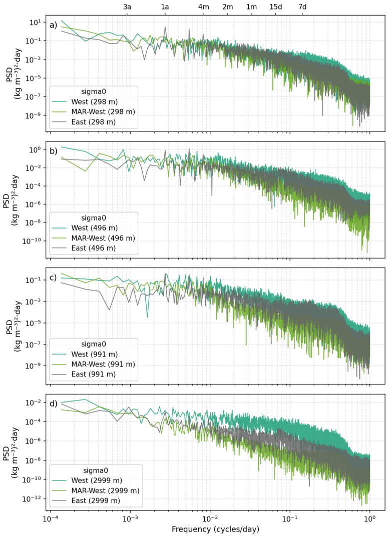
- Target variable: σ_0 (or its anomaly/MTS).
- Four depth levels: 300, 500, 1000, 3000 m.
- Three sites: West, MAR-West, East.
- `np.argmin(np.abs(z_all - target))` finds the nearest actual sensor depth.
- Human-readable period axis added on top via `ax.secondary_xaxis('top')` with fixed ticks at calendar periods.

```
f  = sfft.rfftfreq(N, d=dt)
df = 1.0 / (N * dt)

X   = sfft.rfft(raw_ts)
psd = (2.0 * dt / N) * np.abs(X)**2

ax.loglog(f[1:], psd[1:],
color=site_colors[s_idx],
linewidth=1.0, alpha=0.85,
label=label)

```

Note: index 0 is the DC (mean) component; we skip it to avoid the infinite spike on a log axis.

PSD of σ_0 at Four Depths

- Seasonal peak ($1\,\mathrm{yr}^{-1}$) dominates the shallow panels and weakens with depth. Why?
- Sub-annual energy at 4–6 months persists to depth at the East site: can it be baroclinic Rossby waves?
- West site shows elevated high-frequency energy at depth. Why?
- Run the FFT on the MTS data. Explain it.



The Great Trade-Off: Time *vs* Frequency

- The FFT projects *all* N_t time steps onto each frequency.
- What is gained: exact freq. resolution $\Delta f = 1/(N\Delta t)$, full σ^2 conservation.
- What is lost: the x-axis of the spectrum is frequency, time is gone.
- A 90-day wave packet that passed by in 2008 produces a few bumps at $f \approx 1/90\,\text{day}^{-1}$, but the FFT cannot tell you it happened in 2008.
- Key assumption—ergodicity: the spectrum is the same in every segment of the record. The real ocean *violates* this assumption.
- $\Rightarrow$ We need a transform that retains both time and frequency information.

Non-Stationarity in the Real Ocean

- The most interesting ocean signals are episodic:
  - Mesoscale eddies: 30–120 days, propagate then dissipate.
  - Baroclinic Rossby waves: months–years, arrive episodically from the east.
  - Meddies, ENSO teleconnections, deep convection events.
- They are transient: the dominant period and amplitude *change with time*.
- The FFT of a decade of data averages all these events together.
- We want the scalogram: $P(\omega, t)$—power as a function of freq. and time.

Indulging on Risky Analogies

Analogy: Music

A standard FFT of a symphony tells you which notes were played but not when. A spectroscope resolves frequencies in time. Wavelets are the oceanographer's spectroscope.

Analogy: Movies

Detective movies revolve around finding who has motive, means and *opportunity*. The latter allows one to know when the crime was committed, solving it by finding who has no alibi.

Next Step

We will use the Continuous Wavelet Transform (CWT) to recover the when and the what, to impersonate the maestro or the detective.

Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. **Numerical Intuition**
6. The Continuous Wavelet Transform (CWT)

Hands-on Part 1: Interactive Environment

- Open your terminal and launch `ipython`.
- We will build a "Fakelet" analysis from scratch to understand how signals are decomposed in time-frequency space.

```
import numpy as np
import matplotlib.pyplot as plt
from scipy.signal import oaconvolve
from scipy.signal.windows import blackman

# Setup constants
YEAR = 365.25
dt = 7  % 1-week sampling
t = np.arange(0, int(30 * YEAR), dt)
M = len(t)

```

- We are simulating 30 years of weekly observations.

- We create 5 distinct "oceanic events" with different periods and amplitudes.

```
amplitudes = [1.0, 2.0, 1.5, 2.0, 1.0]
periods = np.array([0.25, 0.33, 0.5, 1, 2]) * YEAR
n_blocks = len(periods)
block_size = M // n_blocks

S = np.zeros(M)
for j, (A, w_j) in enumerate(zip(amplitudes, periods)):
    end_idx = (j + 1) * block_size if j < n_blocks - 1 else M
    idx = slice(j * block_size, end_idx)
    S[idx] = A * np.sin(2 * np.pi / w_j * t[idx])

S += 0.3 * np.random.randn(M) % Add some noise

```

Hands-on Part 3: Time Series Inspection

- Plot the resulting series. Be curious!
- Question for Oceanographers: Is this signal stationary?
- Would a single FFT spectrum accurately describe the physics of the second block versus the fourth?

(Expect a piecewise sine wave with noise)

```
plt.figure(figsize=(10,4))
plt.plot(t, S)
plt.ylabel('Delta Temp (deg C)')
plt.xlabel('Time (days)')
plt.show()
```

Hands-on Part 4: The Intuition of Cascading Filters

- Standard Wavelets use a "Mother Wavelet" stretched at different scales.
- Our "Fakelet" approach uses a cascading bank of Blackman windows.
- **The Core Idea**:
  1. Convolve the signal with a window of size s.
  2. Save that "band" of energy.
  3. Subtract that energy from the signal (*the residual*).
  4. Move to the next scale using only the residual.
- This prevents "double counting" variance across scales.

Hands-on Part 5: Implementing the Filter Bank

```
s1, s2, ns = int(periods.min()/8), int(periods.max()), 30
window_sizes = np.logspace(np.log10(s1), np.log10(s2), ns).astype(int)
scg = np.zeros((ns, M))

residual = S.copy()
# Iterate from largest window to smallest
for i, npf in enumerate(reversed(window_sizes)):
    bla = blackman(npf)
    bla /= bla.sum() % Normalize
    scg[i, :] = oaconvolve(residual, bla, mode='same')
    residual -= scg[i, :] % The cascading step

scg = scg[::-1, :] % Re-order for plotting
frequencies = 0.2 * (window_sizes.astype(float) ** -1)
```

- Be curious, plot some `scg`s just to see what the filters do.

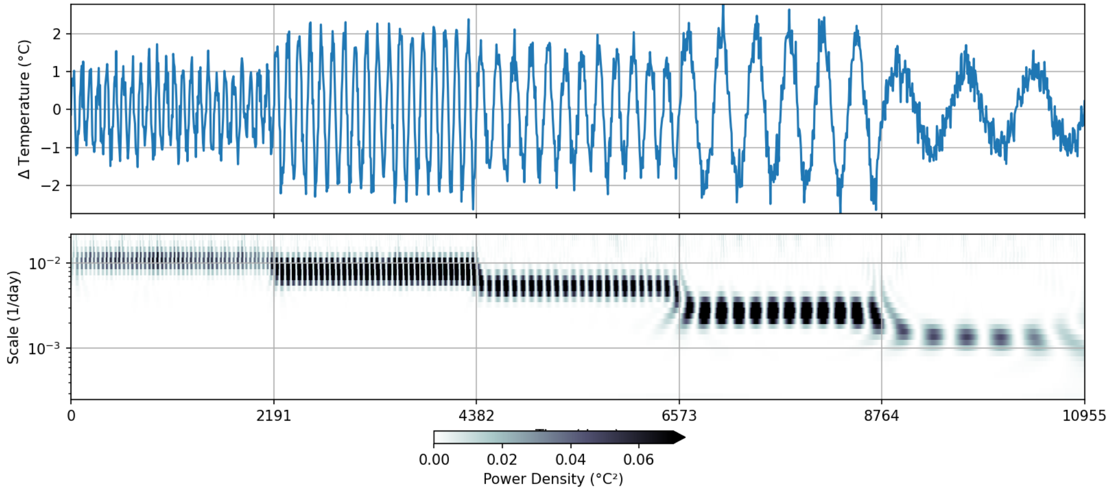
Hands-on Part 6: Visualizing the Pseudo-Scalogram

- Now we map the energy in the Time-Frequency plane.

```
plt.pcolormesh(t, frequencies, scg**2, cmap='bone_r', vmin=0, vmax=0.07)
plt.yscale('log')
plt.ylabel('Scale (1/day)')
plt.xlabel('Time (days)')
plt.colorbar(label='Power Density')
plt.show()

```

- Analysis: Notice how the energy "blobs" align perfectly with the timing of our 5 synthetic blocks.
- This is exactly what the **CWT** will do, but with more mathematical rigor!

Hands-on Part 7: Ta-daa



Outline

1. The RAPID-MOCHA Array
2. Data Exploration
3. Data Processing & Anomaly Decomposition
4. Spectral Analysis via FFT
5. Numerical Intuition
6. The Continuous Wavelet Transform (CWT)

The Continuous Wavelet Transform Definition

- Instead of infinite sines and cosines, we break our signal down using a localized, decaying wave packet called a mother wavelet $\psi(t)$.
- Think of the mother wavelet as a convolution filter function, but with a variable number of points.
- As the shape is the same, but the width varies, the algorythm acts as multiple band-pass filters.

The Continuous Wavelet Transform Definition ... continued

- The CWT ($W_n(s)$) shifts (translates) and scales (dilates) the mother wavelet ψ along our timeline:

$$W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^* \left[\frac{(n' - n)\Delta t}{s} \right]$$

- Where:
  - n is the translation parameter (moving forward across time steps).
  - s is the scale parameter (stretching or squeezing the wavelet to capture low or high frequencies).
  - ψ^* is the complex conjugate of the localized wave function.

Interpreting the Scalogram & The Cone of Influence

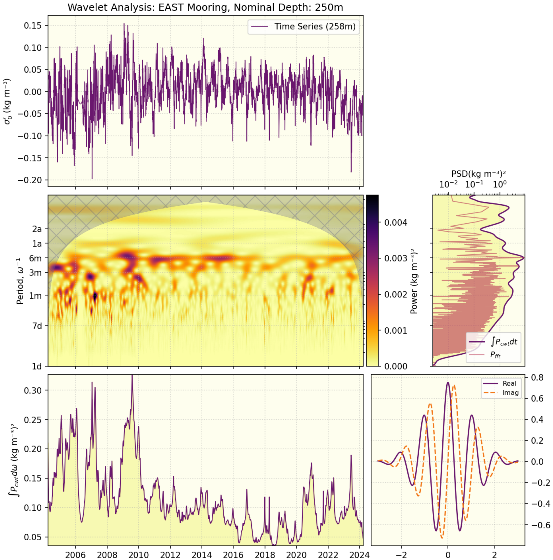
- The local Wavelet Power Spectrum (the scalogram) is defined as $|W_n(s)|^2$.
- Cone of Influence (COI): Because our timeline has boundaries at t_0 and t_N, padding the edges introduces zero-boundary artifacts. The COI marks the region where edge effects drop the power by a factor of e^{-2}.
- Data outside the COI is reliable; data inside is not.
- Integrals: Integrating the scalogram across time yields the global wavelet spectrum (directly comparable to a smoothed FFT PSD).

Program 6: Implementing CWT via PyCWT

- We apply the CWT to the real RAPID-MOCHA σ_0' anomaly vectors.
- We use a standard complex Morlet wavelet to balance localization in both t & ω.
- PyCWT automates scales generation, periods calculation, and construction of the COI boundary matrix.
- Power fields are normalized by the variance of the time-series (σ^2).

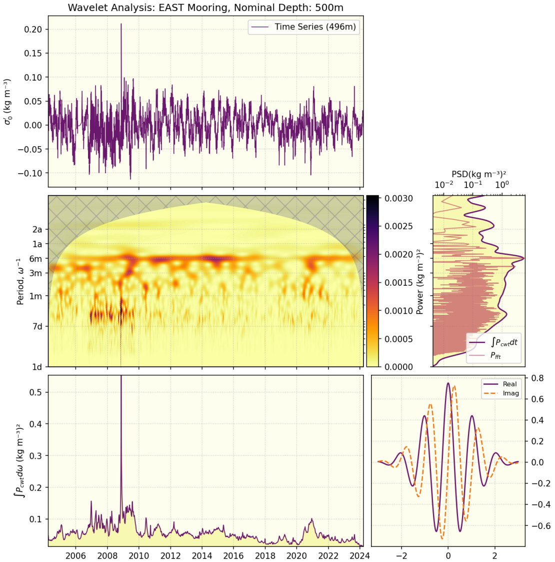
```
import pycwt as wavelet

# Define scale parameters
dj = 1/12  # Sub-octave resolution
s0 = 2 * dt  # Smallest scale
j1 = 7 / dj  # Number of scales

# Perform Continuous Wavelet Transform
wave, scales, freqs, coi, fft, fftfreqs = \
    wavelet.cwt(anomaly_ts, dt, dj, s0, j1,
                wavelet.Morlet())

# Calculate normalized power spectrum
power = (np.abs(wave)) ** 2
real_periods = 1 / freqs
```

- Bottom panel and top right panel are $\int d\omega$ and $\int dt$. Comments?
- No seasonal signature, no trend, just the fun part.
- Localized high-frequency features and distinct energy packets.
- What does the slanted pattern in 2008–2010 means?



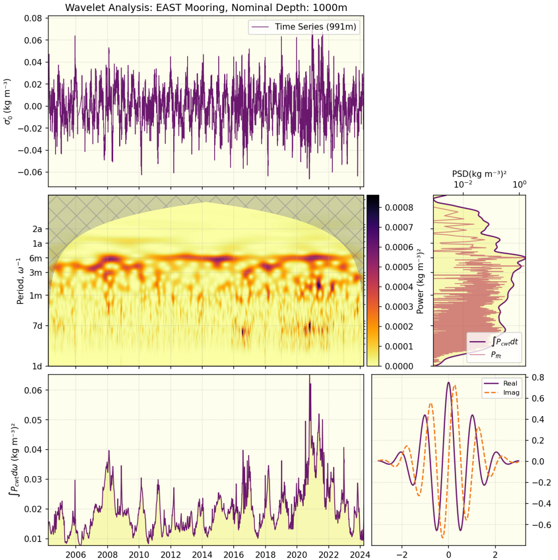
Intermediate Layer (496 m)

- Scalogram profiles reveal strong variance localization at high ω.
- Weird, localized block & spike of high-frequency signals.
- Are semiannual Rossby waves permitted at 26° N? What about a period of 4-months?



Deep Layer (991 m)

- In the NADW layer, all frequencies are damped: compare colorbars.
- The scalogram shows localized activity in 2016 and 2019-2022.
- This is for the anomaly, try with the original data.
- Sooo... what's up, doc?



What Have We Accomplished?

- Isolated the non-deterministic components of the RAPID-MOCHA data.
- Filtered out known signals to enhance the underlying phenomena.
- Identified the core pitfalls of traditional, FFT-based spectral analysis.
- Linked wavelets to convolution filters, building intuition from 1st principles.
- Applied wavelets to real-world data to extract meaningful insights.

♠ You now have the instruments. Go play and have fun!