Gap-filling Landsat with MODIS + AlphaEarth (Daily, Dynamic, Fourier Seasonal Baseline)

This README documents the **daily** gap-filling method implemented in the latest code. It uses:

1) a **seasonal (day-of-year) baseline** learned with a short **Fourier series** (smooth climatology), 2) a **MODIS short-term anomaly** computed via a centered **±7-day window**, and 3) a **year-level bias** predicted from **AlphaEarth** embeddings.

Observed Landsat values are preserved; predictions are used only where Landsat is missing.

Data & Pre-processing

We work band-by-band for the shared Landsat/MODIS bands `blue, green, red, nir08, swir16, swir22`.

Scaling to physical units

Most products are stored as scaled integers. Convert each DataArray `da` using NetCDF attributes:

$$\text{value}_{\text{phys}} = \text{scale\_factor} \cdot \text{value}_{\text{int}} + \text{add\_offset}.$$

(Code: `to_physical`.)

Site-mean time series

For a band b we spatially average over (y, x) to get a 1-D series per sensor:

$$S_b^L(t) = \text{mean}_{y,x}\, L_b(t, y, x), \qquad S_b^M(t) = \text{mean}_{y,x}\, M_b(t, y, x).$$

(Code: `site_series`.)

AlphaEarth yearly descriptor

We compress the (time, 64, y, x) embedding to **one scalar per year** by averaging over `(band,y,x)` and resampling yearly:

$$E_y = \text{mean}_{k\in[64],y,x}\, \text{AE}_k(y, x, \text{year} = y).$$

(Code: `ae_year_series`.)

Seasonal baseline via Fourier (what “seasonal” means)

“Seasonal” = **day-of-year (DOY)** dependence, not a single annual mean. We fit a smooth function $\mu_b(d)$ of DOY $d \in [1, 365.25]$ using a short Fourier series of order H :

$$\mu_b(d) \approx \beta_0 + \sum_{h=1}^{H} \left(\beta_h^{(s)} \sin(2\pi h d/365.25) + \beta_h^{(c)} \cos(2\pi h d/365.25)\right).$$

We fit μ_b^L from S_b^L and μ_b^M from S_b^M by **least squares** on available (irregular) timestamps, then evaluate on **every day** of the target index. (Code: `seasonal_mu_fourier(series, target_idx, order=H)`.)

- **Order** H controls smoothness. Typical: 2–4. Higher H captures sharper seasonality but risks overfitting.

MODIS short-term aggregation (dynamic alignment)

Instead of forcing a fixed 14-day grid, we keep true timestamps and compute a centered daily window around any time t :

$$S_b^{M\,\mathrm{win}}(t) = \frac{1}{|W(t)|} \sum_{\tau \in W(t)} S_b^M(\tau), \quad W(t) = \{\tau : |\tau - t| \le w \text{ days}\}.$$

We use $w = 7$ by default (a ±7-day rolling mean). (Code: `modis_window_daily`.)

Fitting parameters on overlaps

Let $\mathcal{O}$ be Landsat observation times. Denote DOY by $d(t)$ and year by $y(t)$.

Anomalies

$$A_b^L(t) = S_b^L(t) - \mu_b^L(d(t)), \qquad A_b^M(t) = S_b^{M\,\mathrm{win}}(t) - \mu_b^M(d(t)).$$

Short-term coupling

Estimate the slope k_b that maps MODIS anomalies to Landsat anomalies (OLS through the origin):

$$\boxed{k_b = \frac{\sum_{t \in \mathcal{O}} A_b^L(t)\, A_b^M(t)}{\sum_{t \in \mathcal{O}} \left(A_b^M(t)\right)^2}}$$

Year-level residual modeled by AlphaEarth

Residual per year after removing MODIS anomaly effect:

$$\bar{r}_{b,y} = \mathrm{mean}_{t \in \mathcal{O}_y} \left(A_b^L(t) - k_b A_b^M(t) \right).$$

Regress this on the AlphaEarth descriptor (years with overlap):

$$\boxed{\bar{r}_{b,y} \approx a_b + \gamma_b E_y}$$

so a_b (intercept) and γ_b (slope) are fitted by simple linear regression of $\bar{r}_{b,y}$ on E_y .

Daily prediction rule

For **every day** t on the target daily index:

$$\boxed{\hat{S}_b^L(t) = \mu_b^L(d(t)) + k_b \big(S_b^{M\,\mathrm{win}}(t) - \mu_b^M(d(t)) \big) + \big(a_b + \gamma_b E_{y(t)} \big)}$$

Observed values are kept:

$$S_b^{L,\mathrm{filled}}(t) = \Big\{ S_b^L(t), \quad t \in \mathcal{O},\ \hat{S}_b^L(t),\ \text{otherwise}.$$

- First term: **Landsat seasonal baseline** (smooth Fourier curve of DOY).
- Second term: **MODIS short-term anomaly**, scaled by k_b .
- Third term: **AlphaEarth year bias** shared by all days in that year.

Mapping to the code

- `to_physical` — apply `scale_factor` / `add_offset`.
- `site_series` — spatial mean per band → 1-D Series.
- `ae_year_series` — scalar E_y per year.
- `seasonal_mu_fourier` — fits and evaluates μ_b^L , μ_b^M on a daily index.
- `modis_window_daily` — computes $S_b^{M\,\mathrm{win}}(t)$ (±7d rolling mean).
- `fill_one_band_daily` — fits k_b, a_b, γ_b using overlaps and produces **daily** filled series; returns labels and parameters for plotting.

Choosing hyperparameters

- **Fourier order** H : start at 3. Lower if seasonality is very smooth; raise cautiously if curves are underfit.
- **Window half-width** w : default 7 days. Smaller tracks faster changes; larger smooths noise.

Plotting & labeling

- **Y-axis**: use each band's metadata after scaling: `long_name (units)`. Surface reflectance is typically unitless in [0,1]`). - **Per-band figures**: title format *"Gap-filled Landsat — DAILY — band: `<band>` [k=..., a=..., γ=...]"* --- ## Extensions - **Season-varying coupling:** estimate \(k_{b,h} per season (e.g., by month or DOY bin) instead of one global k_b .
- **Vector AE:** regress on the 64-D AE vector with ridge: $\bar{r}_{b,y} \approx a_b + \boldsymbol{\gamma}_b^\top \mathbf{e}_y$.
- **Quality control:** mask cloudy/low-quality Landsat before fitting; optionally clip predictions to observed per-season quantiles.
- **Derived indices:** compute NDVI, etc., from `nir08` & `red` (after scaling) and apply the same pipeline.

Glossary of symbols

- b — spectral band (e.g., `nir08`).
- t — day; $y(t)$ — year of day t ; $d(t)$ — day-of-year of day t .
- $S_b^L(t)$ — site-mean Landsat for band b at time t .
- $S_b^M(t)$ — site-mean MODIS (daily).
- $S_b^{M\,\mathrm{win}}(t)$ — centered $\pm w$ -day mean of MODIS around t .
- $\mu_b^L(d), \mu_b^M(d)$ — Fourier seasonal baselines for Landsat and MODIS.
- $A_b^L(t), A_b^M(t)$ — anomalies relative to seasonal baselines.
- k_b — OLS slope mapping MODIS anomalies to Landsat anomalies.
- $\bar{r}_{b,y}$ — year-mean residual after MODIS scaling.
- E_y — AlphaEarth yearly descriptor.
- a_b, γ_b — intercept and slope from yearly residual regression.
- $\hat{S}_b^L(t)$ — predicted Landsat used to fill gaps (daily).