



Forecasting the intermittent solar resource

Cyril Voyant, Research Director — Mines Paris – PSL, Centre O.I.E., Sophia Antipolis

UE14 — Mines Paris engineering curriculum (English version of the 2026 course)

Applied research project: modelling and forecasting global solar irradiance

MINES PARIS – PSL

CENTRE O.I.E.

SOPHIA ANTIPOLIS

Cite as: Voyant, C. (2026). Forecasting the intermittent solar resource (UE14 course, English version). Mines Paris – PSL. Licence CC BY 4.0.

Web companion with the worked example: cyrilvoyant.github.io/teaching/solar-forecasting-basics.html

Course outline

01

Why forecast solar power?

Grid issues, units, the Ajaccio data

02

Solar radiation

Direct, diffuse, atmosphere, Sun position

03

Time series and forecasting

Components, horizon, a first AR(1) model

04

Clear sky and clear-sky index

Solis model, CSI, stationarity

05

AR model and least squares

Building X and y, least squares, choosing p

06

Evaluation, ELM and online learning

Persistence, metrics, ELM, sliding window, frugality

Appendices: wide-sense stationarity, conditional expectation, AIC/BIC criteria. Glossary and references at the end.

Project objectives

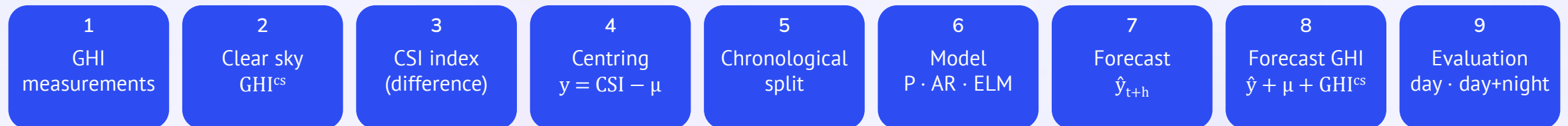
Context

- Solar irradiance is **intermittent** (clouds).
- The power grid must anticipate photovoltaic production.
- Need: **reliable short-term** forecasts (here 1 h).

By the end, you will be able to

- compute a **clear-sky index** (CSI);
- build and estimate an **AR(p)** model by least squares;
- build an **ELM** and measure its variability;
- compare models with **persistence** (MAE, RMSE, skill score);
- update a model **online** and judge its **frugality**.

The full chain we will build



This chain comes back in every chapter; the current step is highlighted.

Notation (used throughout the course)

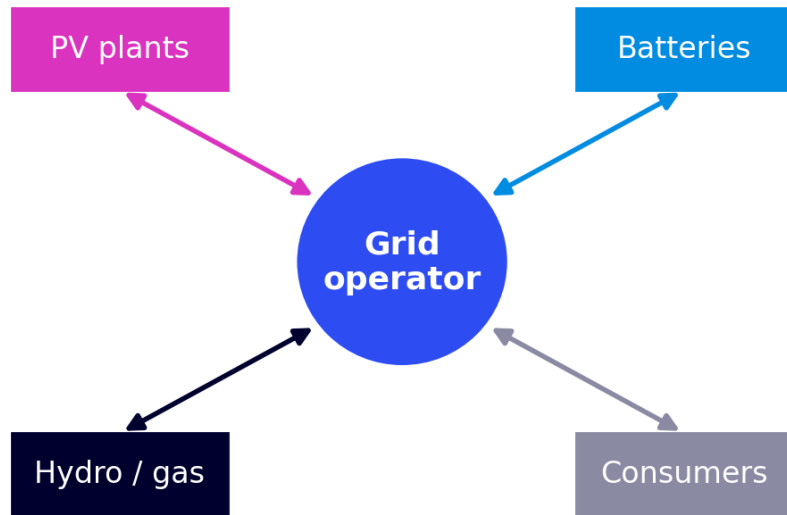
Symbol	Meaning
GHI_t	global horizontal irradiance measured at hour t (Wh/m^2)
GHI_t^{cs}	clear-sky GHI (Solis model)
CSI_t	clear-sky index: $GHI_t - GHI_t^{cs}$ (Wh/m^2)
y_t	centred CSI, the modelled series: $y_t = CSI_t - \mu$
$\hat{y}_{t+h}$	forecast of y made at t for $t + h$
h	horizon, in number of steps (1 step = 1 h)
p	number of lags (past values used)
GHI_{t+h}^{fc}	$\hat{y}_{t+h} + \mu + GHI_{t+h}^{cs}$; 0 at night or if negative
e_t	forecast error: $GHI_t^{fc} - GHI_t$
μ	mean CSI (training period)

Symbol	Meaning
$\mathbf{X}, \mathbf{y}$	input matrix, target vector
$\mathbf{w}$	AR model coefficients
ε_t	error (unpredictable part)
$\mathbf{W}_{in}, \mathbf{b}$	random ELM weights and biases
$\mathbf{H}, \boldsymbol{\beta}, L$	hidden outputs, output weights, number of neurons
$\theta_z, \delta, \omega, \varphi$	zenith angle, declination, hour angle, latitude
$\alpha; W$	ridge regularization; sliding-window length
N	number of examples (rows of $\mathbf{X}$)
Δt	time step (here 1 h)

Convention: **bold lower case** = vector, **bold upper case** = matrix, plain = number. Time is always a subscript: y_t, y_{t-1}, CSI_{t+h} .

Why forecast? Grid and solar production

A 1-hour PV forecast lets the operator act before clouds arrive



The grid operator balances PV, storage, other plants and consumers.

Solar output varies with clouds. To integrate it into the grid, it must be **anticipated**.

Grid stability

Anticipate production drops to keep the frequency stable.

Storage

Charge or discharge batteries at the right time.

Supply / demand balance

Adjust the other plants (hydro, pumped storage*, gas...).

Economic value

Sell production at the best price and cut operating costs.

* Pumped-storage hydropower: a reversible dam that stores energy by pumping water uphill.

Irradiance, irradiation and photovoltaic output

Irradiance (W/m²)

Power received per m² at a given instant. Example: 1,000 W/m² in full sun at noon.

Irradiation (Wh/m²)

Energy received per m² over a duration. Over one hour: hourly irradiation in Wh/m² = mean irradiance in W/m².

Our data

Hourly GHI in Wh/m². Météo-France files are in J/cm²: 1 J/cm² = 2.78 Wh/m², so Wh/m² = J/cm² / 0.36.

From irradiance to electrical power

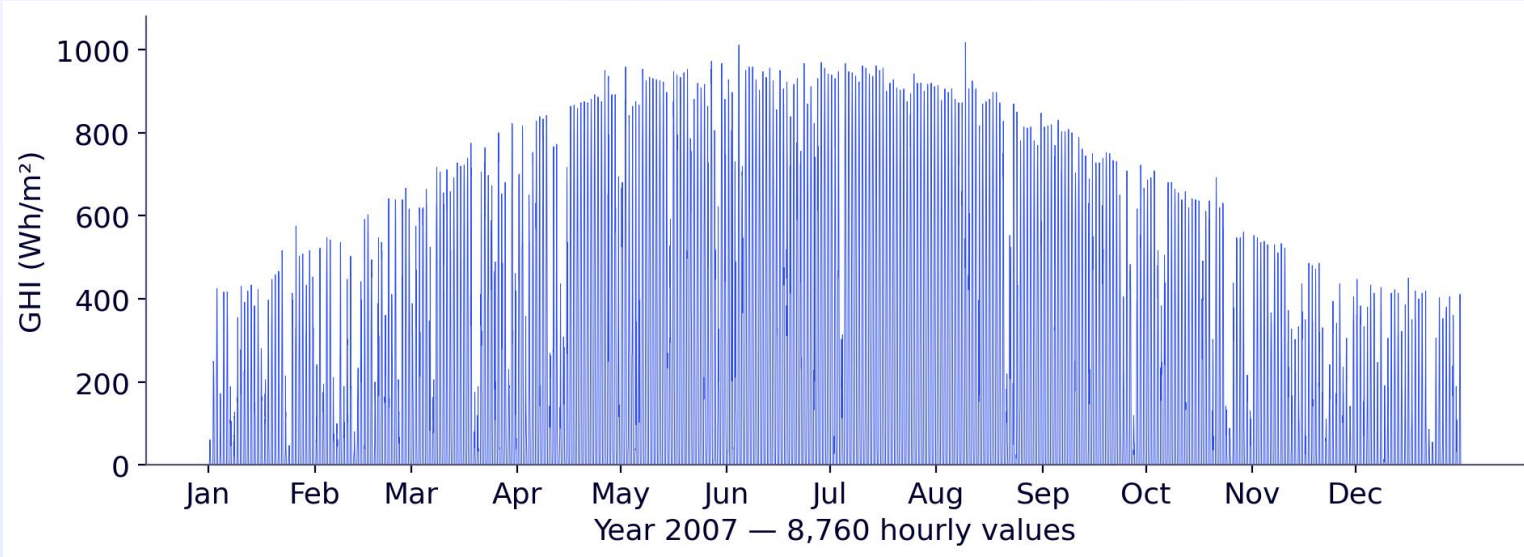
$$P \approx \eta \cdot S \cdot G$$

η : module efficiency ($\approx 20\%$), S : area (m²), G : irradiance on the module (W/m²).

Example: 10 m² under 800 W/m² $\rightarrow P \approx 0.20 \times 10 \times 800 = 1.6$ kW.

Output is proportional to irradiance: **forecasting GHI is forecasting production.**

The dataset: Ajaccio, 1998–2007

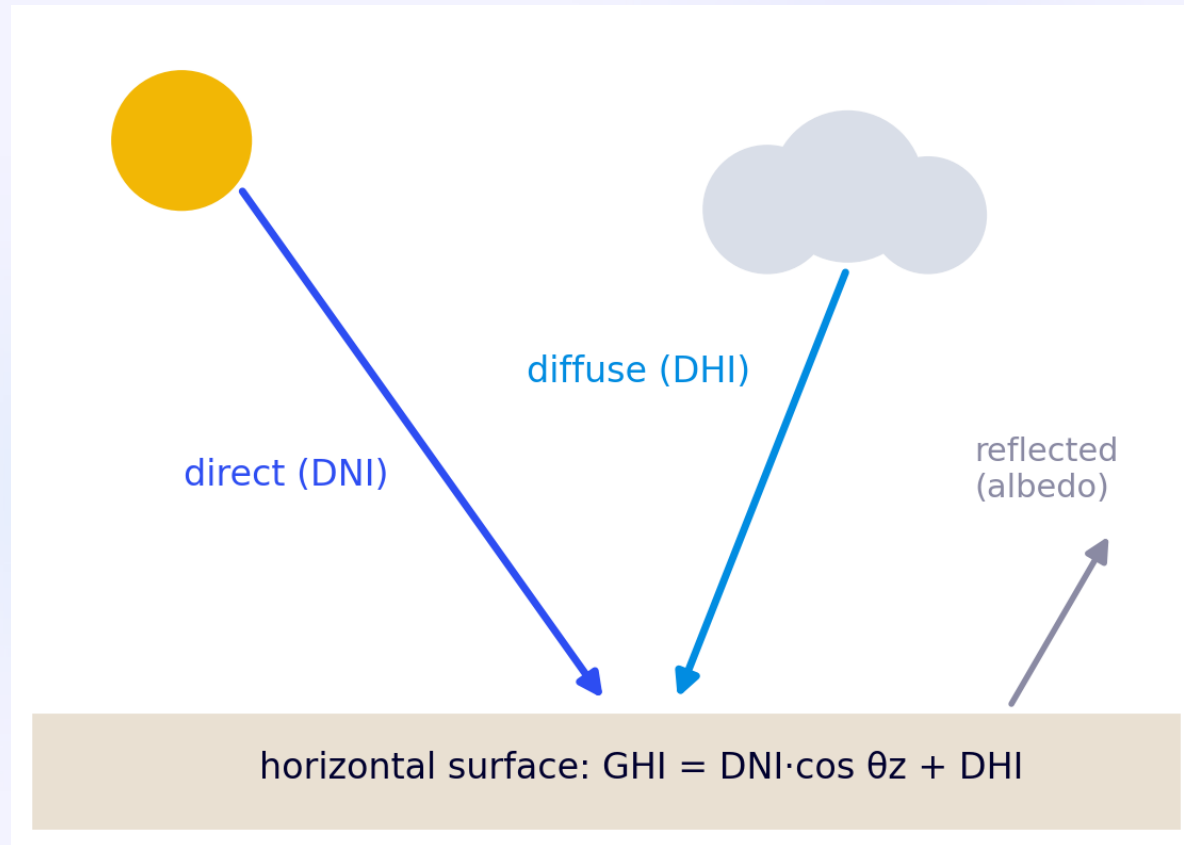


Hourly GHI measured at Ajaccio in 2007 (Météo-France).

- **Site:** Ajaccio (Corsica), 41.9° N, 8.8° E, 4 m.
- **Variable:** hourly GHI (Wh/m^2).
- **Period:** 1998–2007, i.e. 87,600 hours.
- **Maximum:** 1,025 Wh/m^2 .
- **Daytime mean:** 368 Wh/m^2 .

Ten years of measurements: 8 years to **learn**, 1 year to **tune** the model, 1 year to **test** it (chapter 5).

Components of global radiation



θ_z : zenith angle (between the Sun and the vertical). GHI = *Global Horizontal Irradiance*.

Direct (DNI)

Received straight from the Sun, measured on a surface perpendicular to the rays (*direct normal*). Dominant under clear sky.

Diffuse (DHI)

Scattered by air, aerosols and clouds. Dominant under overcast sky.

Reflected (albedo)

Sent back by the ground. Counts only for a **tilted surface**, not for GHI.

Crossing the atmosphere

Through the atmosphere, radiation is attenuated by four mechanisms.

Absorption

Water vapour, ozone, CO₂: part of the energy is absorbed.

Rayleigh scattering

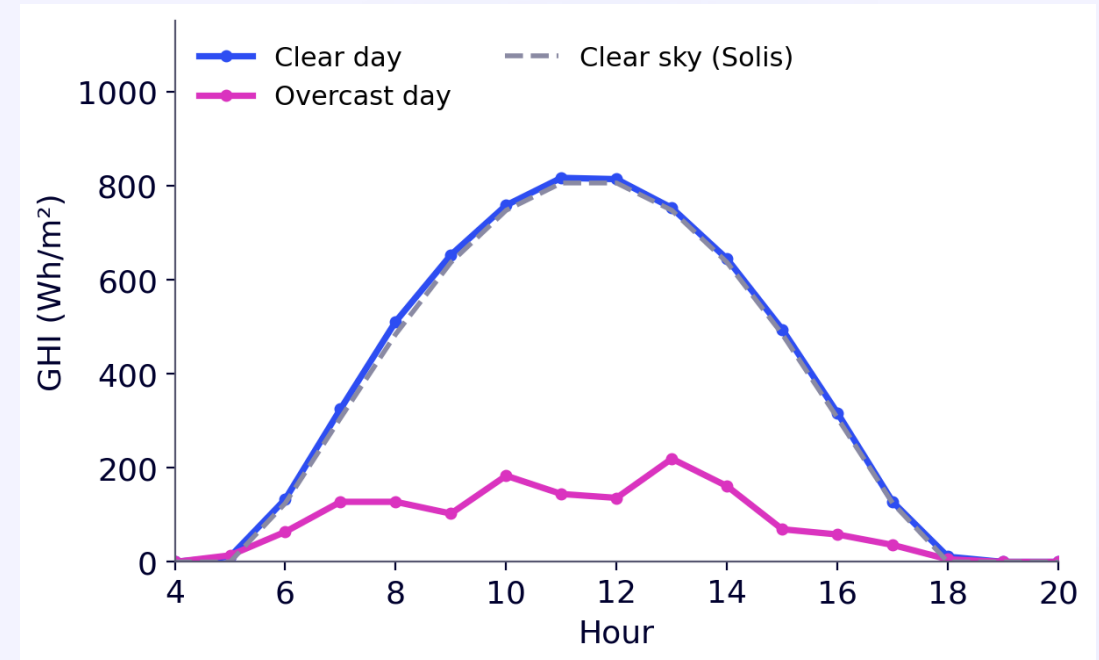
By air molecules: it makes the sky blue.

Mie scattering

By aerosols (dust, sea salt, pollution): depends on turbidity.

Clouds

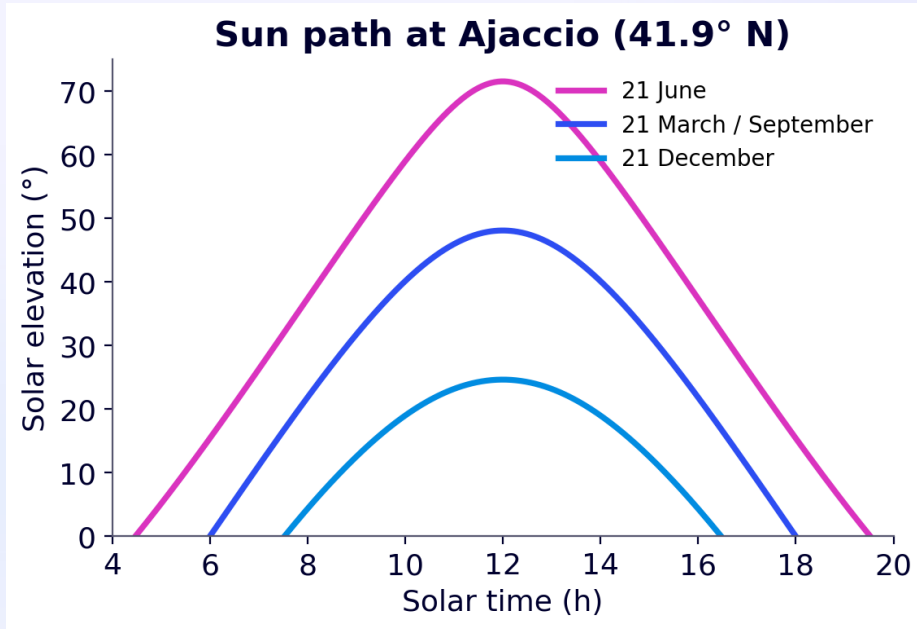
Reflection, scattering and absorption: **main cause of intermittency**.



Ajaccio, 2007: a clear day and an overcast day.

Under overcast sky, GHI drops here to **about 20 %** of its clear-sky value.

The position of the Sun



Solar elevation over the day at Ajaccio, by season.

- **Zenith angle** θ_z : between the Sun and the vertical. The larger it is, the thicker the atmosphere crossed.
- **Declination** δ : angle between the Sun's direction and the equatorial plane; from -23.45° to $+23.45^\circ$ over the year.
- **Hour angle** ω : 15° per hour, zero at solar noon. **Latitude** φ : 41.9° at Ajaccio.

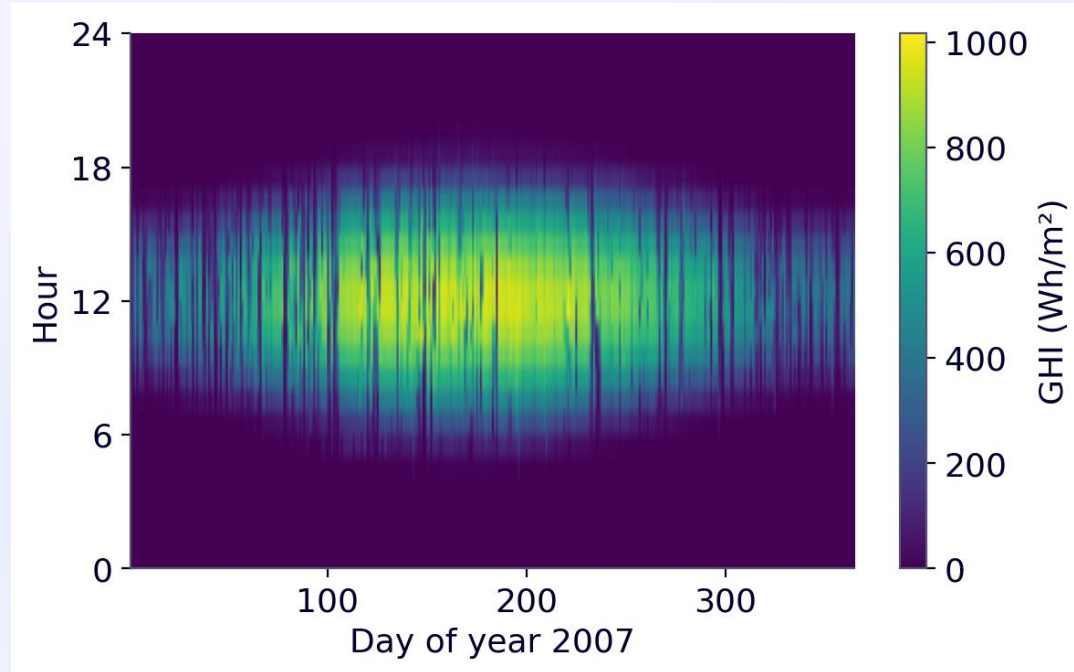
$$\cos \theta_z = \sin \varphi \cdot \sin \delta + \cos \varphi \cdot \cos \delta \cdot \cos \omega$$

$$G_0 = I_0 \cdot (d_0/d)^2 \cdot \cos \theta_z$$

G_0 : extraterrestrial radiation; $I_0 \approx 1361 \text{ W/m}^2$ (solar constant; 1367 in older codes).

The Earth–Sun distance d varies by $\pm 1.7\%$, so $(d_0/d)^2$ varies by $\pm 3.3\%$.

The double seasonality of GHI



Hourly GHI in 2007: hours of the day $\times$ days of the year (Ajaccio).

Daily seasonality

Zero at night, maximum around solar noon: a **24-h** cycle well described by the clear-sky model.

Annual seasonality

The amplitude of the daily cycle changes with the season (declination, day length): a **365-day** cycle.

Raw GHI is **not stationary**: its mean depends on the hour and the month. Both cycles must be removed before modelling.

What is a time series?

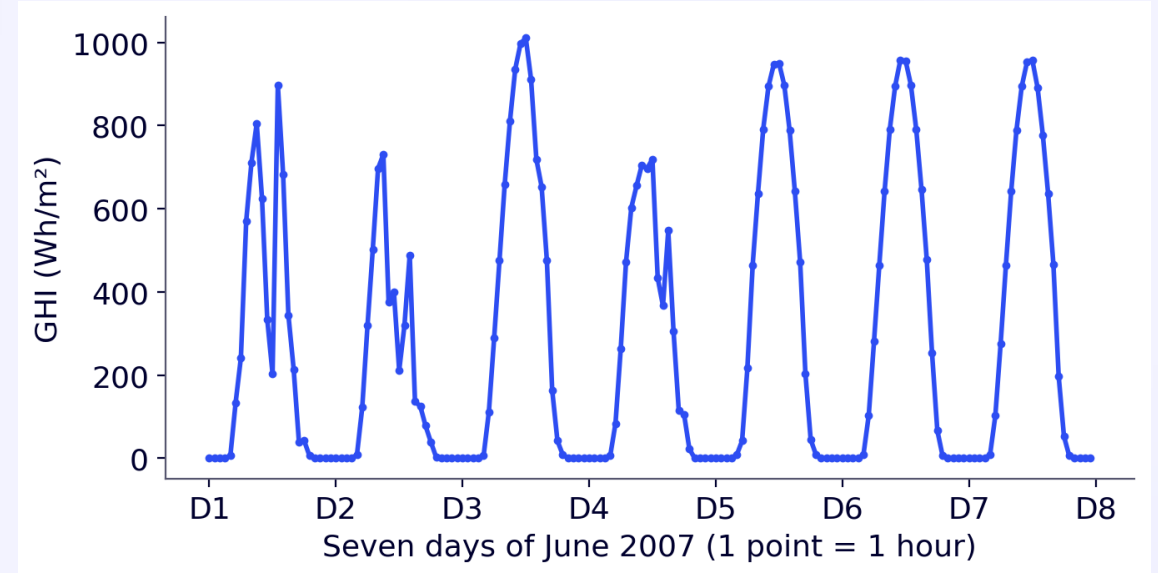
A sequence of measurements $y_1, y_2, \dots, y_N$ of one quantity, taken at a **constant time step** Δt . Here $\Delta t = 1$ h.

Three properties

- **Regularity**: measurements are equally spaced.
- **Dependence**: the past informs the future.
- **Randomness**: each value has an unpredictable part.

What we want for modelling

- **Stationarity**: stable mean and variance over time.
- **Complete series**: no missing values.



Ajaccio: one week of hourly GHI.

Raw GHI is **not** stationary (day/night and seasonal cycles): that is the whole point of chapter 4.

Components of a time series

Trend T_t

Slow long-term change: increasing, decreasing or none.

Seasonality S_t

Regular, repeatable cycles (day, year). Very strong for GHI.

Cycle C_t

Long, not strictly periodic oscillations (climate phenomena).

Residual R_t

Unexplained random part. Ideally **white noise** (independent values, zero mean).

For GHI: seasonality dominates, with two cycles (daily and annual).

Decomposing a time series

Additive model

$$y_t = T_t + S_t + C_t + R_t$$

When the amplitude of fluctuations stays constant.

Multiplicative model

$$y_t = T_t \times S_t \times C_t \times R_t$$

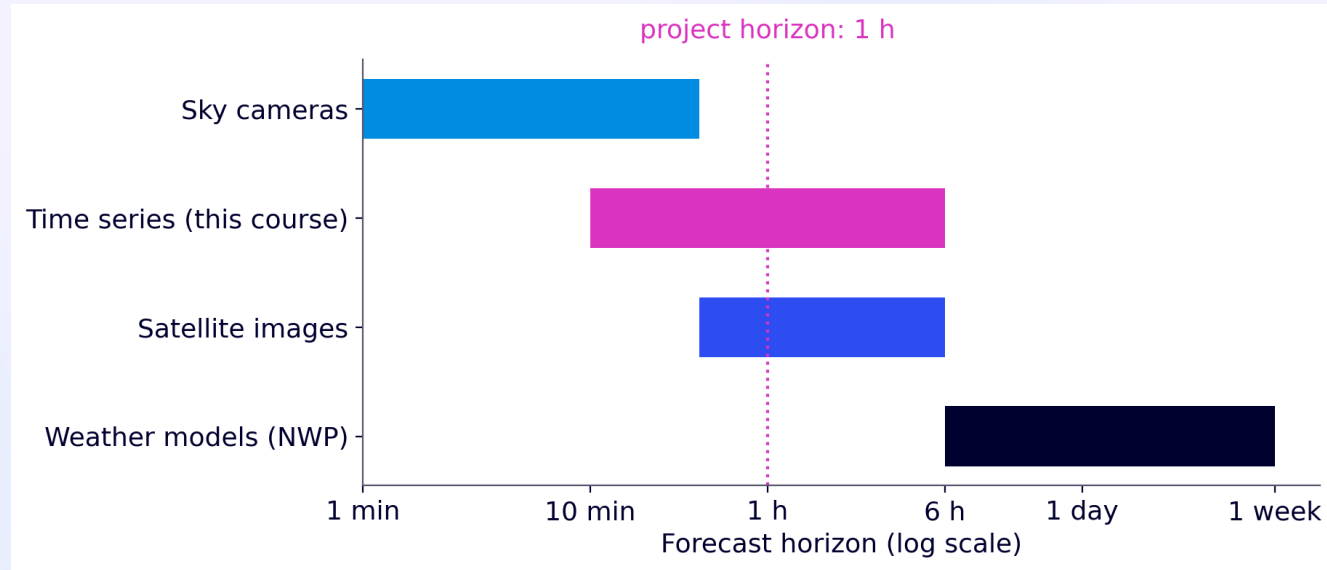
When the amplitude grows with the level (GHI: small in winter, large in summer).

Why decompose?

- Remove trend and seasonality (the part **predictable without effort**);
- keep a residual that is easier to model;
- check that this residual looks like white noise.

Link with what follows: subtracting clear-sky GHI, $CSI_t = GHI_t - GHI_t^{cs}$, applies the **additive model**: both cycles go at once. The ratio (multiplicative model) also exists: chapter 4.

Horizon, time step and types of forecast



Usual methods by horizon (orders of magnitude).

- **Time step** $\Delta t = 1 \text{ h}$; **horizon** h in steps: $h = 1 \rightarrow 1 \text{ h}$, $h = 3 \rightarrow 3 \text{ h}$.
- **Direct forecast**: one model trained for each horizon h .
- **Recursive forecast**: forecast $t + 1$, then feed that forecast back to forecast $t + 2$, and so on.

In this course: 1-h horizon, direct forecast.

Forecasting: a predictable part and an error

$$y_{t+1} = f(y_t, y_{t-1}, \dots, y_{t-p+1}) + \varepsilon_{t+1}$$

The predictable part

f sums up what the past tells us about the future. The forecast is $\hat{y}_{t+1} = f(y_t, \dots, y_{t-p+1})$.

The error

$\varepsilon_{t+1} = y_{t+1} - \hat{y}_{t+1}$ is the unpredictable part. If the model is good, ε looks like white noise.

The goal

Choose f so that errors are **as small as possible** in mean square.

The framework suits GHI: regular measurements and strong hour-to-hour dependence (cloud inertia). Further reading: appendix B.

A first example: the AR(1) model, with numbers

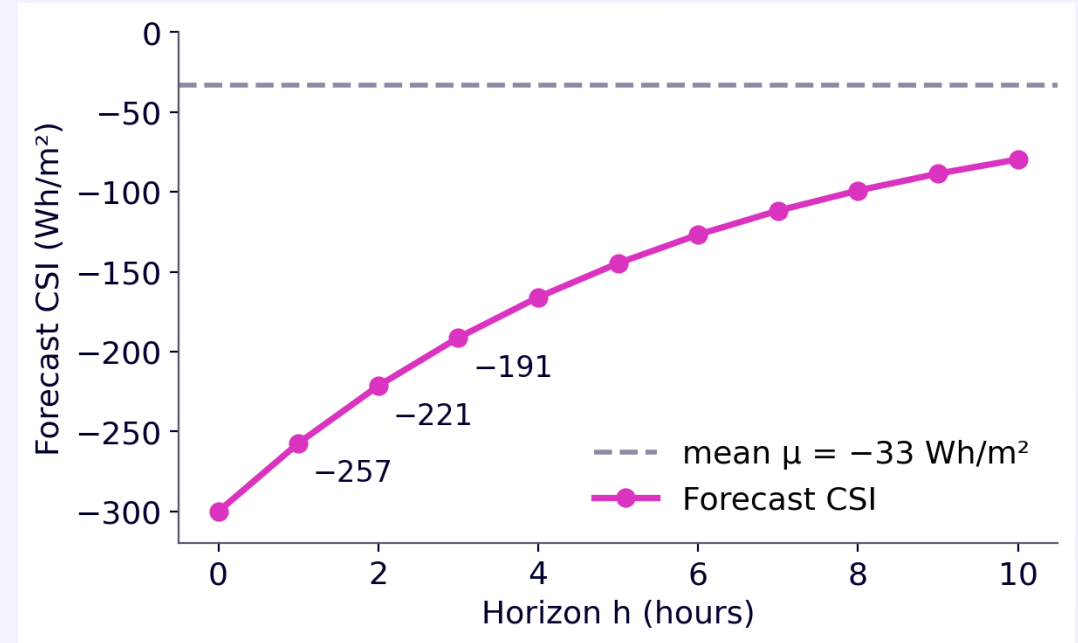
$$y_{t+1} = w \cdot y_t + \varepsilon_{t+1}$$

- $y_t = \text{CSI}_t - \mu$: gap of the CSI to its mean; w : weight of the last gap.
- $|w| < 1$: the series always returns to its mean.

$$\hat{y}_{t+h} = w^h \cdot y_t$$

Ajaccio: $\mu \approx -33 \text{ Wh/m}^2$ and $w \approx 0.84$. A cloud: $\text{CSI}_t = -300 \text{ Wh/m}^2$, so $y_t = -267$.

$\hat{y}_{t+1} = 0.84 \times (-267) = -224$: forecast CSI = -257 Wh/m^2 ; then -221 and -191 at 2 h and 3 h.



The forecast CSI returns to its mean as h grows.

The clear-sky model

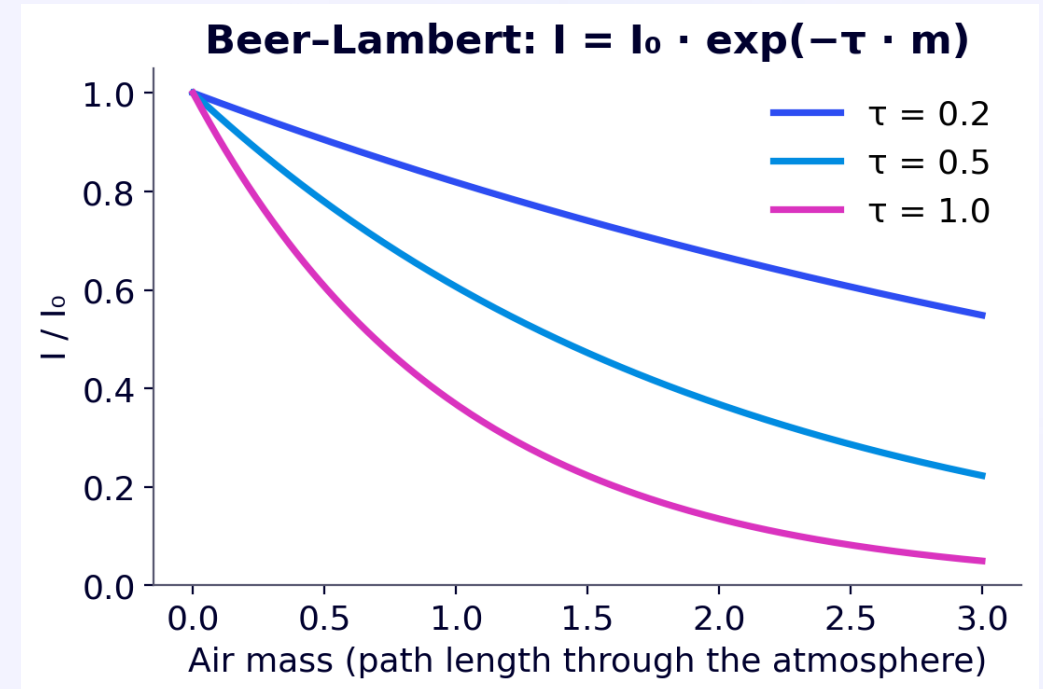
What is it?

The GHI one would measure **without clouds**, given the Sun's position and a clear atmosphere.

Inputs

- site position (latitude, longitude, altitude);
- Sun position (θ_z);
- water vapour and aerosols (AOD: aerosol optical depth).

Role: GHI^{cs}_t is the reference used to remove the day/night and seasonal cycles.



Beer-Lambert law: intensity decays exponentially with the path length crossed.

The Solis model

Solis (Ineichen, 2008) is a simplified version of a radiative-transfer model. It starts from the Beer–Lambert law:

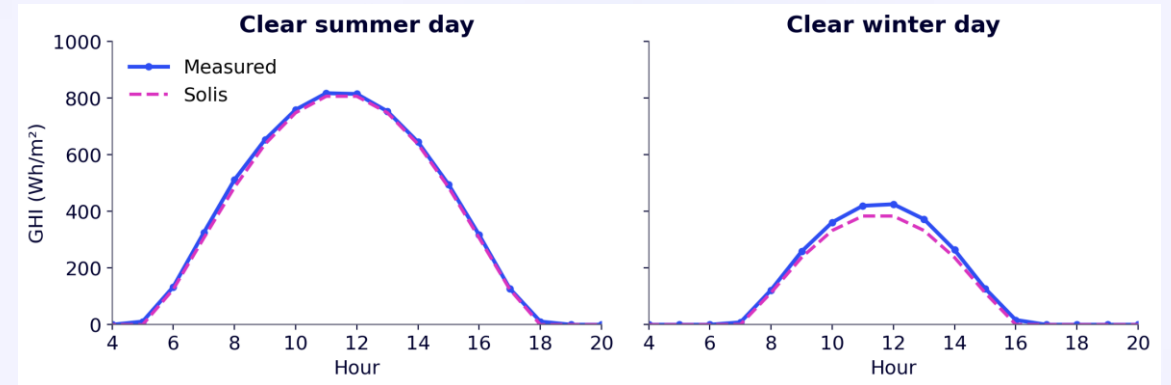
$$\text{GHI}^{\text{cs}} \approx I_0 \cdot \exp(-\tau / \cos \theta_z) \cdot \cos \theta_z$$

- τ : optical depth (air, water vapour, aerosols);
- Solis adjusts τ and the exponent for water vapour, aerosols and altitude.

Settings: rural zone, water column 3.79 cm, AOD 0.16, $I_0 = 1367 \text{ W/m}^2$.

The hourly value is the **mean of the past hour**: the Sun is computed **at mid-hour**, $\omega = (i - 12.5) \times 15^\circ$.

Mean clear-sky error at Ajaccio: $\approx 32 \text{ Wh/m}^2$.



Ajaccio, 2007: Solis and measurement under clear sky.

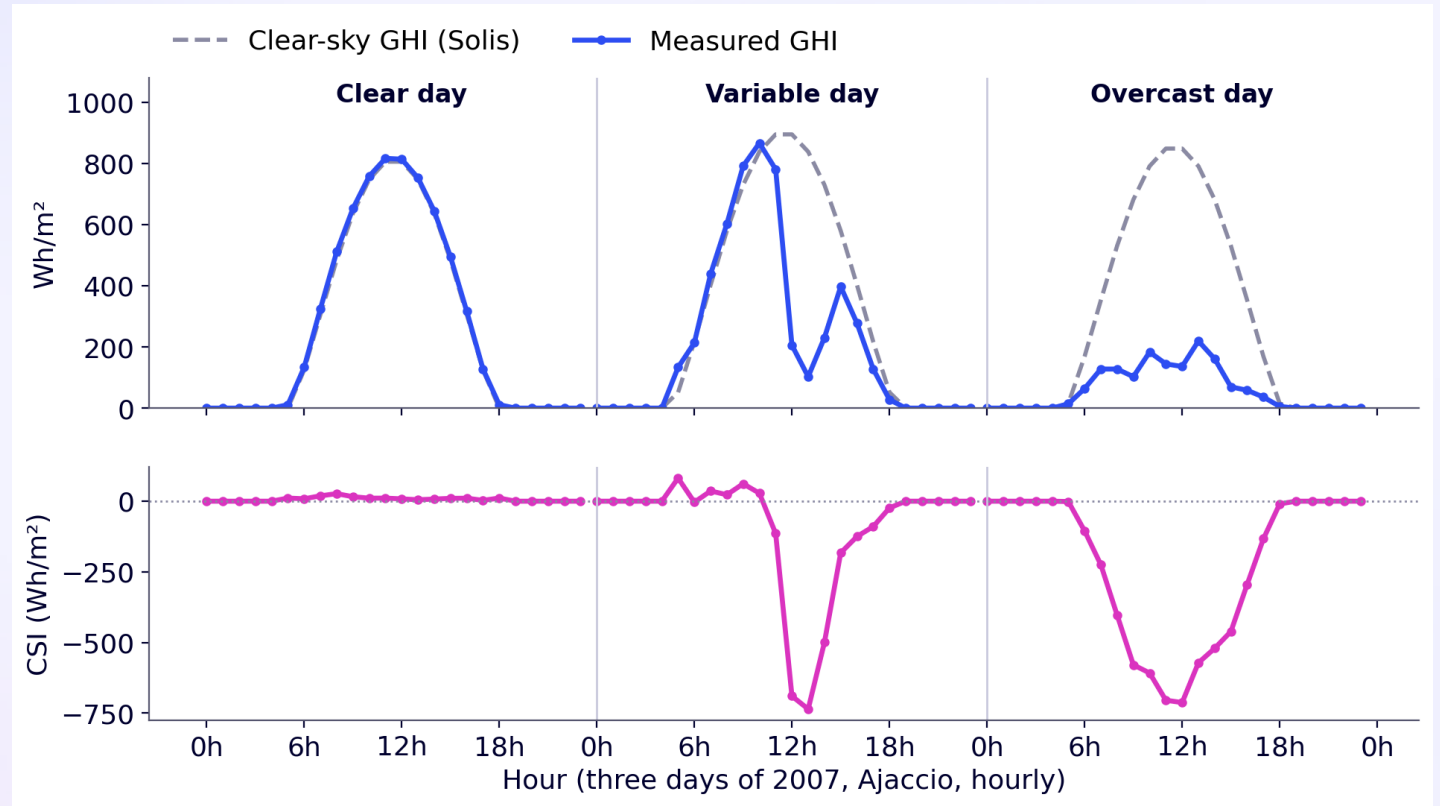
Stationarizing: the clear-sky index (CSI)

$$CSI_t = GHI_t - GHI_t^{cs}$$

- ≈ 0 under clear sky, and **0 at night**;
- **negative** under clouds (down to $-GHI_t^{cs}$);
- mean $\mu \approx -33 \text{ Wh/m}^2$: we **centre** it, $y_t = CSI_t - \mu$.

The procedure

1. compute GHI_t^{cs} , then the CSI;
2. centre: $y_t = CSI_t - \mu$;
3. forecast $\hat{y}_{t+h}$;
4. rebuild $GHI_{t+h}^{fc} = \hat{y}_{t+h} + \mu + GHI_{t+h}^{cs}$ (0 at night, never negative).



Ajaccio, 2007: top, GHI; bottom, CSI. The daily cycle has gone from the CSI.

Two ways to stationarize

The Solis model removes the cycles in two ways:

Difference (additive model) — this course

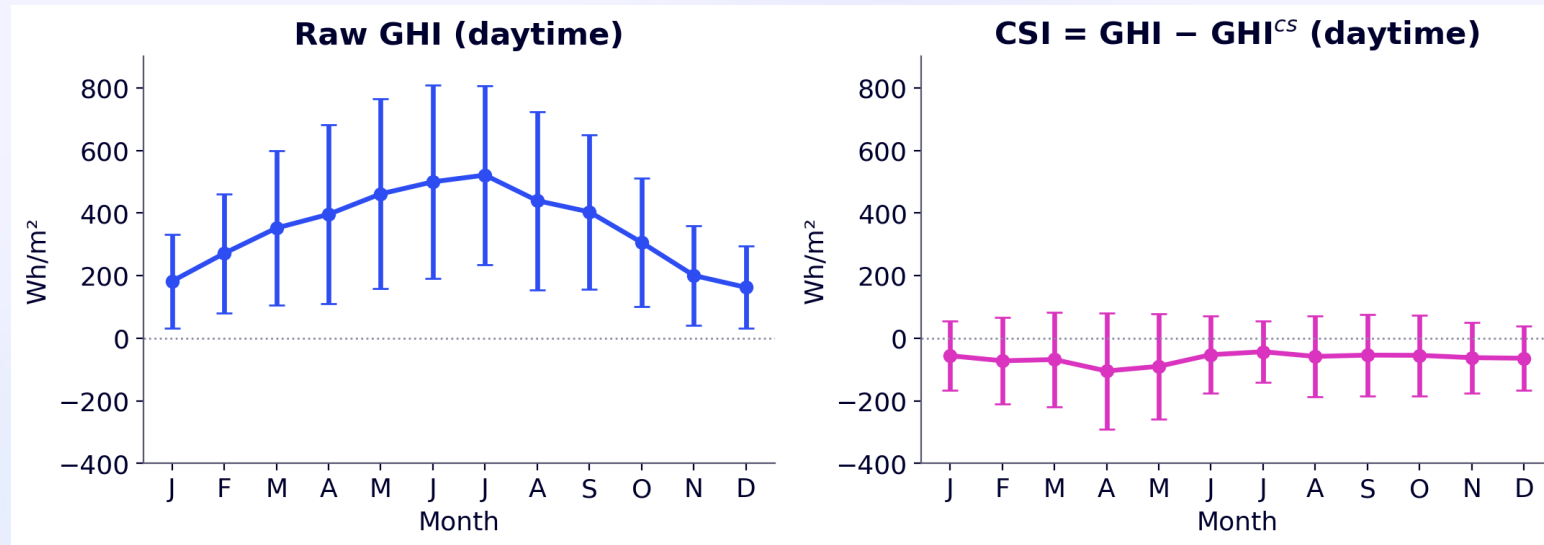
- $CSI_t = GHI_t - GHI_t^{cs}$ in Wh/m^2 ;
- no division: valid **day and night** (at night, $0 - 0 = 0$);
- all hours are kept, no filtering.

Ratio (multiplicative model)

- $k_t = GHI_t / GHI_t^{cs}$, unitless (≈ 1 under clear sky); widely used in the literature;
- at night: **0 / 0**, undefined; at sunrise and sunset, $GHI^{cs} \approx 0$: values explode;
- requires **filtering on solar elevation** (for instance keep $h > 10^\circ$).

For hourly data with nights, the ratio is **poorly suited**: we use the **difference**. The ratio is also very sensitive to small timestamp offsets and to the filtering thresholds chosen (Voyant et al., 2025a; 2026).

Is the CSI stationary?



Mean $\pm$ standard deviation by month, daytime hours, Ajaccio 1998–2007.

- **GHI**: the mean ranges from 163 to 521 Wh/m² depending on the month.
- **CSI**: the mean stays between -105 and -43 Wh/m², close to 0.
- The CSI standard deviation still varies (97 to 185 Wh/m²): more clouds in spring.

The CSI is **approximately stationary**: much more than GHI, though not perfectly. Formal definition: appendix A.

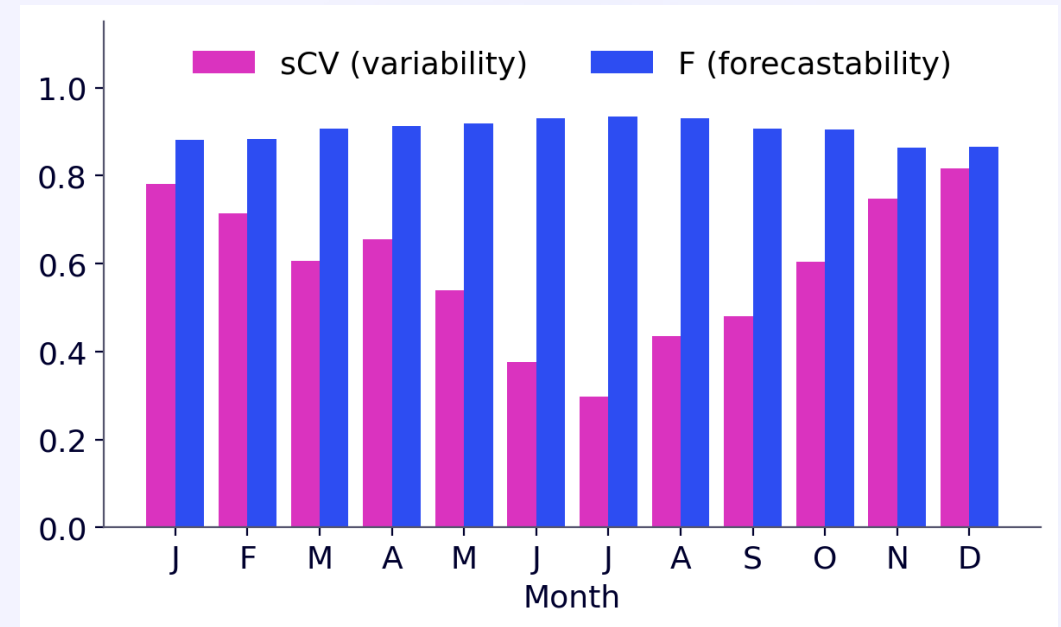
Site variability: the sCV

How far does GHI depart, on average, from clear sky? We measure the gap $\varepsilon_t = \text{GHI}_t - \text{GHI}^{\text{cs}}_t$ (the CSI):

$$\text{sCV} = (\sqrt{2} / (1 - d)) \cdot \sqrt{[\sum_t (\text{GHI}_t - \text{GHI}^{\text{cs}}_t)^2 / \sum_t (\text{GHI}^{\text{cs}}_t)^2]}$$

- **Numerator:** typical size of the gaps to clear sky (the clouds).
- **Denominator:** typical size of clear-sky GHI.
- $\sqrt{2} / (1 - d)$ scales the worst case to 1: clear sky, then very overcast, every other hour ($d = 0.2$: even under overcast sky, at least 20 % of clear sky remains).

sCV = 0: always clear; sCV = 1: maximum variability. **It does not depend on the horizon:** it is a property of the site.



Ajaccio 1998–2007: sCV and F at 1 h, month by month (Voyant et al., 2026).

Ajaccio: sCV = 0.52 over 1998–2007.

Winter: 0.8 (changeable sky); July: 0.3 (often clear sky).

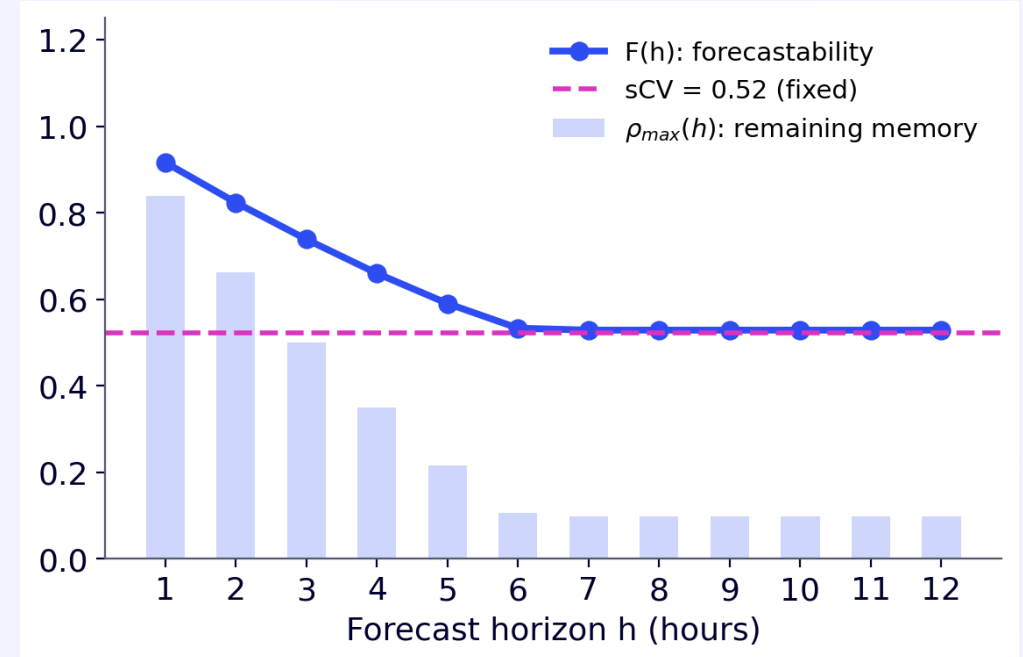
Forecastability: F depends on the horizon

A variable series is not necessarily unpredictable: if the gaps to clear sky **last**, they can be anticipated.

- $\rho(k)$: correlation between the clear-sky gap now, ε_t , and k hours later, ε_{t+k} (the **memory** of clouds).
- $\rho_{\max}(h)$: the strongest of these correlations for k between h and 12 h: the memory left when forecasting at horizon h .

$$F(h) = 1 - \text{sCV} \cdot (1 - \rho_{\max}(h))$$

- perfect memory ($\rho_{\max} = 1$): $F = 1$, everything is predictable;
- no memory ($\rho_{\max} = 0$): $F = 1 - \text{sCV}$.

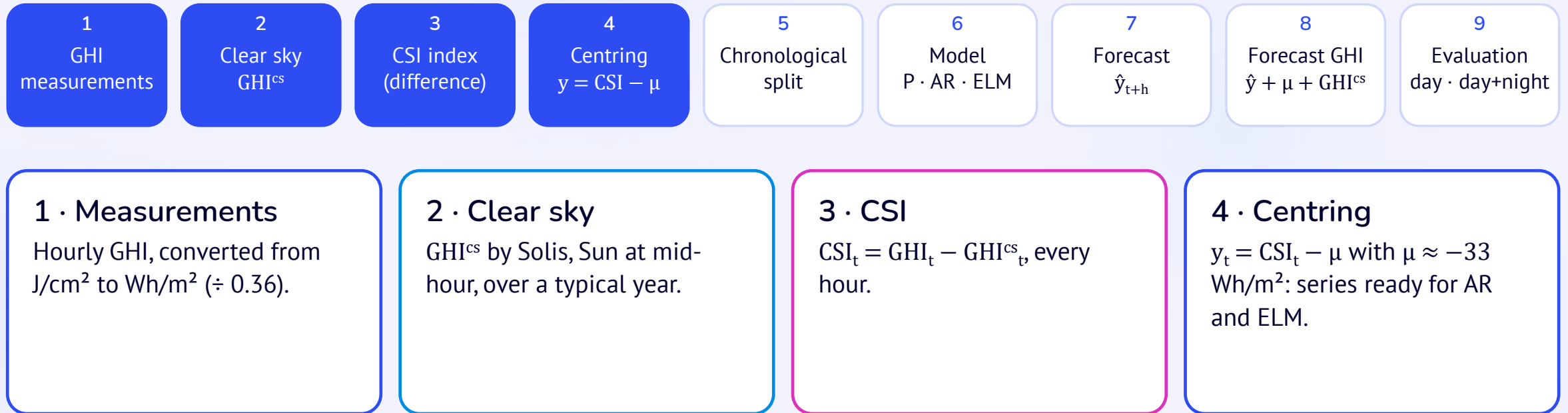


Ajaccio 1998–2007: F and remaining memory by horizon (Voyant et al., 2026).

$F = 0.92$ at 1 h, 0.74 at 3 h, 0.53 at 6 h.

The atmosphere is **chaotic**: cloud memory fades within a few hours. sCV does not move.

Pre-processing step by step



With the difference, nights cause no problem ($0 - 0 = 0$): **no filtering on solar elevation**, unlike the ratio.

From the AR model to linear regression

An **AR(p)** model forecasts the next value from the last p values:

$$y_{t+1} = w_1 y_t + w_2 y_{t-1} + \dots + w_p y_{t-p+1} + \varepsilon_{t+1}$$

Matrix form

$$\mathbf{y} = \mathbf{X} \mathbf{w} + \boldsymbol{\varepsilon}$$

- $\mathbf{y}$ ($N \times 1$): the values to forecast;
- $\mathbf{X}$ ($N \times p$): one row per example, with the p lags;
- $\mathbf{w}$ ($p \times 1$): the coefficients to estimate;
- $\boldsymbol{\varepsilon}$ ($N \times 1$): the errors.

It is a **multiple linear regression**. Since y is centred, **no intercept is needed**.

Building $\mathbf{X}$ and $\mathbf{y}$: a worked example

Five centred CSI values (Wh/m^2): $y_1 = -10$; $y_2 = -25$; $y_3 = -240$; $y_4 = -180$; $y_5 = -30$. Take $p = 2$.

Row	y_t	y_{t-1}	Target y_{t+1}
$t = 2$	-25	-10	-240
$t = 3$	-240	-25	-180
$t = 4$	-180	-240	-30

- The two middle columns form $\mathbf{X}$ (3×2), the last one forms $\mathbf{y}$ (3×1).
- Each row shifts the previous one by one step.

In practice: all hours are kept, nights included (at night, $y = -\mu \approx 33$). At Ajaccio: 8 years, about 70,000 rows for training.

Ordinary least squares (OLS)

We look for the $\mathbf{w}$ that minimizes the sum of squared errors:

$$J(\mathbf{w}) = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = \sum_i (y_i - \mathbf{x}_i^T \mathbf{w})^2$$

At the minimum, the derivative is zero:

$$-2 \mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) = 0 \Rightarrow \mathbf{X}^T\mathbf{X} \mathbf{w} = \mathbf{X}^T\mathbf{y}$$

- Conditions: many more examples than coefficients ($N \gg p$);
- columns of $\mathbf{X}$ not proportional to each other.

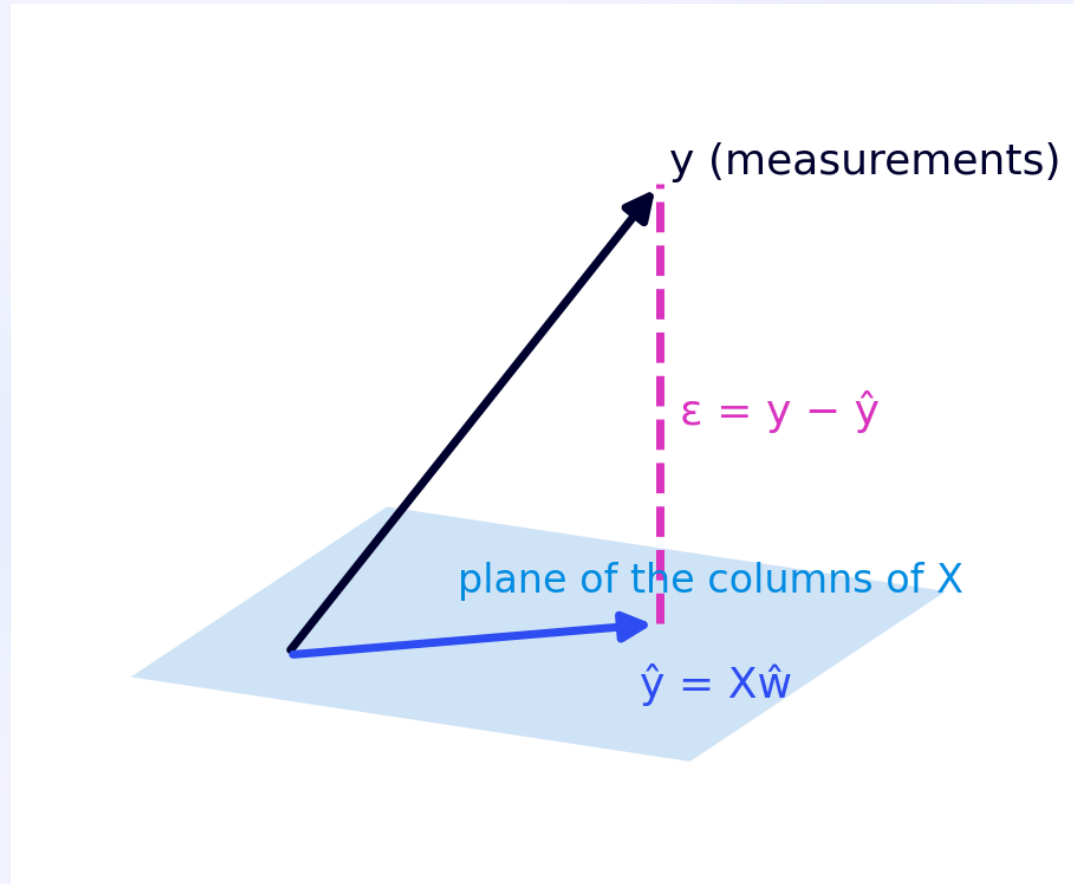
$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{X}^T\mathbf{y}$$

Exact solution, in one step: no iteration, a single minimum.

Ajaccio, AR(2): $\hat{\mathbf{w}} = (0.97; -0.15)$, computed in 5 ms.

The last gap y_t weighs by far the most.

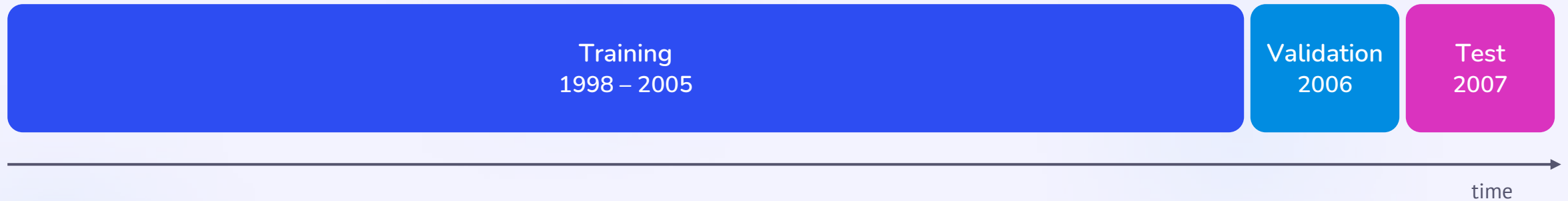
Geometric interpretation



- $X\hat{w}$ is the point of the “plane of the columns of X” **closest** to y : its orthogonal projection.
- The error $\epsilon = y - X\hat{w}$ is **perpendicular** to this plane.
- Pythagoras: $\|y\|^2 = \|\hat{y}\|^2 + \|\epsilon\|^2$. This is the basis of the R^2 coefficient.

OLS does not “guess” anything: it projects the measurements onto what the inputs can explain.

Training, validation, test



Training

Coefficients are estimated ($\mathbf{w}$, $\boldsymbol{\beta}$).

Validation

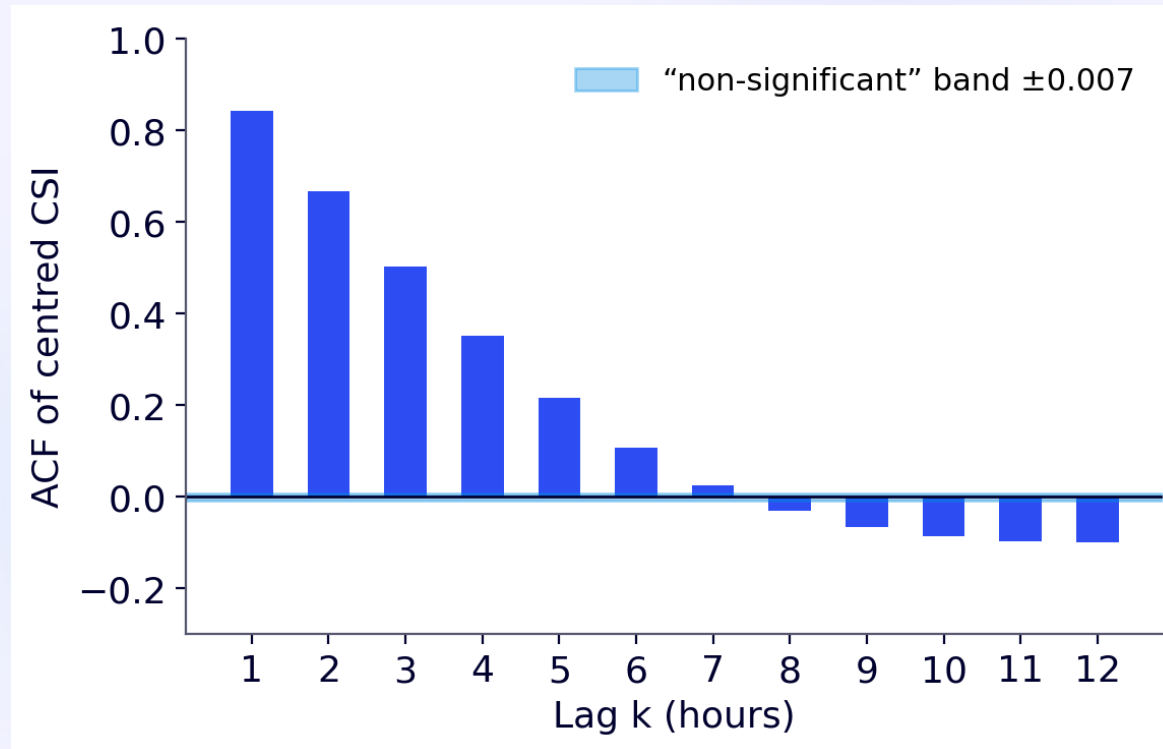
Hyperparameters (p , L) are chosen without touching the test.

Test

Performance is measured **once**, on data never seen.

Chronological split, never shuffled: the past is not forecast with the future. After validation, re-estimate on training + validation.

Choosing p (1/3): autocorrelation (ACF)

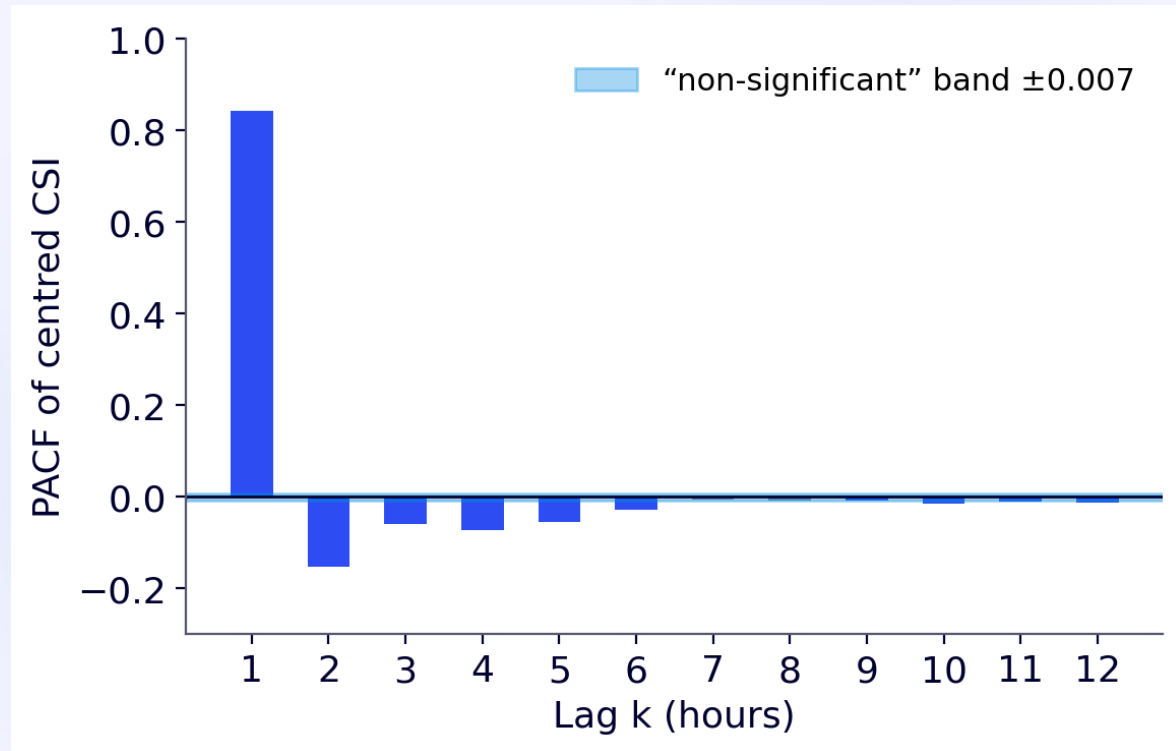


ACF of the centred CSI, Ajaccio 1998–2005, all hours.

- **ACF(k)**: correlation between y_t and y_{t-k} , the value k hours earlier.
- Here $ACF(1) = 0.84$, $ACF(2) = 0.67$, $ACF(3) = 0.50$: it decays slowly.
- But this correlation **propagates step by step**: if y_t depends on y_{t-1} , which depends on y_{t-2} , then y_t and y_{t-2} are correlated even without a direct link.

The ACF therefore does not tell how many lags are **really useful**. The PACF is needed.

Choosing p (2/3): partial autocorrelation (PACF)



PACF of the centred CSI, Ajaccio 1998–2005, all hours.

- **PACF(k)**: once $y_{t-1}, \dots, y_{t-k+1}$ are known, what does y_{t-k} **add**? It is the last coefficient w_k of an $AR(k)$.
- A bar **outside the blue band**: this lag brings direct information.
- For a true $AR(p)$, the PACF vanishes after lag p .

Here $PACF(1) = \mathbf{0.84}$, $PACF(2) = -0.15$, then values below 0.08: the PACF suggests **$p \approx 2$** .

But with $N \approx 70,000$ points, the ± 0.007 threshold is tiny: almost everything is “significant”. Significant $\neq$ useful.

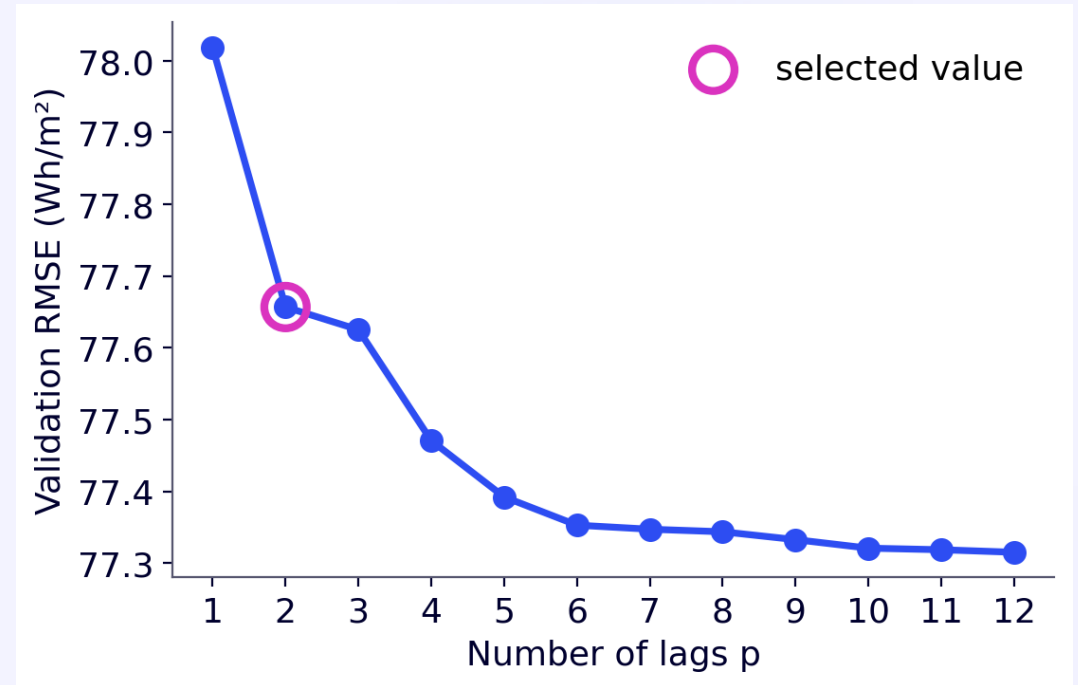
Choosing p (3/3): filter or wrapper approach

Filter

- look at **the data alone**, before any model: PACF, AIC / BIC criteria (appendix C);
- fast, but indirect.

Wrapper

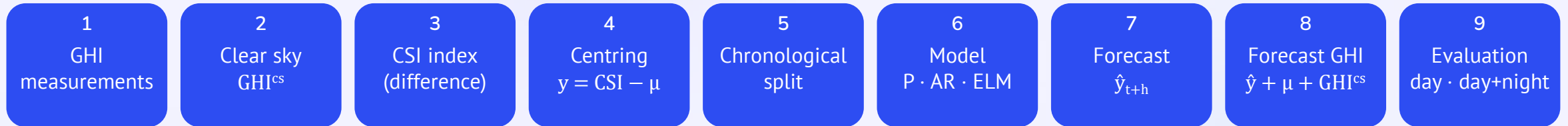
- **train** an $AR(p)$ for $p = 1$ to 12 on 1998–2005;
- measure the **forecast error** on validation (2006);
- driven by the real objective, but more costly.



Wrapper approach: validation RMSE (2006, daytime hours) versus p .

$p = 1$: 78.0; $p = 2$: 77.7; $p = 12$: 77.3 Wh/m². **Parsimony**: the smallest p within 0.5 % of the best, i.e. **$p = 2$** .

Summary: forecasting GHI with an AR(p)



1. Compute GHI^{cs} (Solis), then $\text{CSI}_t = \text{GHI}_t - \text{GHI}^{\text{cs}}_t$; centre it: $y_t = \text{CSI}_t - \mu$.
2. Split: training / validation / test.
3. Choose p on validation.
4. Build $\mathbf{X}$ and $\mathbf{y}$; estimate $\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.
5. Forecast $\hat{y}_{t+1} = \hat{\mathbf{w}}^T \mathbf{x}_t$ with $\mathbf{x}_t = (y_t, \dots, y_{t-p+1})$.
6. Rebuild $\text{GHI}^{\text{fc}}_{t+1} = \hat{y}_{t+1} + \mu + \text{GHI}^{\text{cs}}_{t+1}$; set **0 at night** (Sun below the horizon) and **0 if the forecast is negative**.
7. Evaluate on 2007: daytime, then day + night.

Simple, fast, interpretable: it is the **linear reference** to beat.

Reference models

Every model must be compared with a simple reference. If it does not do better, it adds nothing.

Simple persistence

“It will be as it is now.”

$$GHI_{t+1}^{fc} = GHI_t$$

Ignores the Sun's course.

Clear sky

“It will be sunny.”

$$GHI_{t+1}^{fc} = GHI_{t+1}^{cs}$$

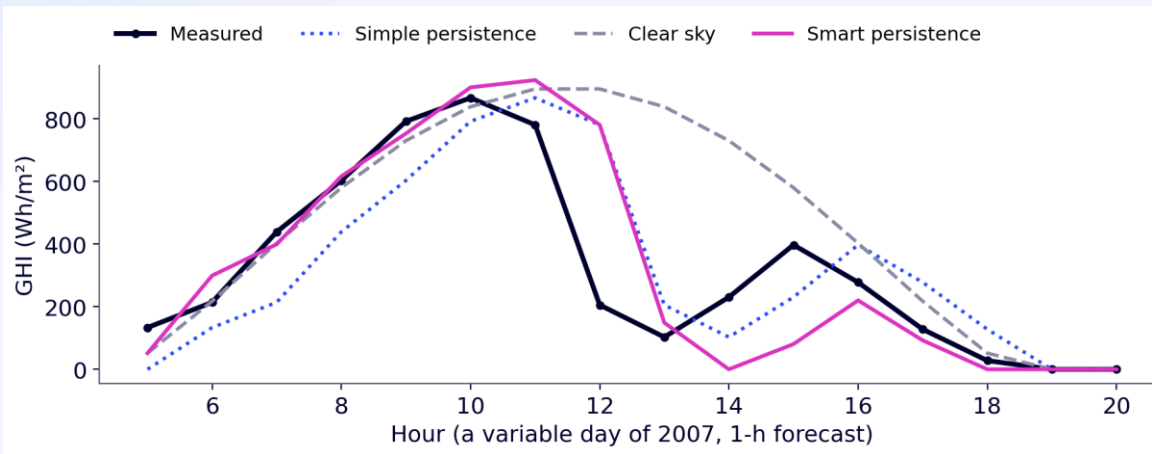
Ignores clouds.

Smart persistence

“The sky will stay the same.”

$$GHI_{t+1}^{fc} = GHI_t - GHI_t^{cs} + GHI_{t+1}^{cs}$$

Keeps the same gap to clear sky.



Smart persistence = AR(1) with $w = 1$: **the reference of the skill score SS.**

Two rules, for every model: 1) at night (Sun below the horizon, $\theta_z > 90^\circ$), forecast = 0; 2) a negative forecast is replaced by 0.

Measuring the error

$$e_t = \text{GHI}_t^{\text{fc}} - \text{GHI}_t \quad (\text{forecast error})$$

$$\text{nMBE} = [(1/N) \sum e_t] / \text{mean}(\text{GHI}_t)$$

$$\text{MAE} = (1/N) \sum |e_t|$$

$$\text{RMSE} = \sqrt{[(1/N) \sum e_t^2]} \quad ; \quad \text{nRMSE} = \text{RMSE} / \text{mean}(\text{GHI}_t)$$

- **nMBE** (bias): > 0 if the model **over-estimates** on average, < 0 if it **under-estimates**. Errors of opposite sign cancel: it is **not** a score for ranking models.
- **MAE**: mean error, in Wh/m^2 , easy to interpret.
- **RMSE**: penalizes large errors more (sudden cloud). Always $\geq$ MAE.
- **nRMSE**: RMSE as % of mean GHI, to compare sites.

Essential check: if $|\text{nMBE}| > 1 \%$, there is probably a problem, or a very poor model. The bias can be removed, provided this is tested; we do not do it here.

Errors are computed on the **rebuilt GHI**, over the **test** year (2007): first on **daytime hours**, then **day + night**. Always say which!

Comparing with persistence: SS and NICE Σ

$$SS = 1 - RMSE_{\text{model}} / RMSE_{\text{smart persistence}}$$

- **Skill score:** > 0 if smart persistence is beaten, < 0 otherwise.
- Looks only at the RMSE.

$$NICE^k = L^k(\text{model}) / L^k(\text{simple persistence})$$

$$L^k = ((1/N) \sum |e_t|^k)^{1/k}; \quad NICE^\Sigma = (NICE^1 + NICE^2 + NICE^3) / 3$$

$e_t = GHI_t^{\text{fc}} - GHI_t$: the forecast error (previous slide).

- $k = 1$: average errors (MAE); $k = 2$: RMSE; $k = 3$: large errors;
- **$NICE^\Sigma < 1$** : the model beats simple persistence on all three norms at once.

Two different references: SS compares with **smart** persistence, $NICE^\Sigma$ with **simple** persistence (Voyant et al., 2025b).

Results at Ajaccio: 1-hour forecast

Model	RMSE day	nRMSE day	nRMSE day + night	nMBE day	SS	NICE ^Σ
Simple persistence	135	36.4 %	51.5 %	-0.9 %	-0.66	1.00
Clear sky	148	39.9 %	56.4 %	+16.4 %	-0.82	1.06
Smart persistence	82	22.0 %	31.0 %	+1.3 %	0.00	0.59
AR(2)	79	21.3 %	30.1 %	+2.0 %	+0.03	0.58
ELM (L = 50, best draw)	76	20.5 %	28.9 %	+1.1 %	+0.07	0.55

Test 2007. Daytime: 4380 hours, mean GHI 372 Wh/m². Day + night: 8759 hours, mean GHI 186 Wh/m². RMSE in Wh/m²; SS and NICE computed on daytime hours.

Smart persistence divides the RMSE of simple persistence by 1.7. AR gains **3 %** over it, ELM **7 %**.

Daytime only, or day + night?

79 → 56

AR RMSE (Wh/m²)

372 → 186

Mean GHI (Wh/m²)

21 → 30 %

AR nRMSE

- At night (Sun below the horizon), every model forecasts 0 by construction and is right: adding these hours **lowers the RMSE** ($\approx \div \sqrt{2}$).
- But the mean GHI is **halved**: the nRMSE **increases**.
- The ranking of models does not change.

An error figure without the list of hours used means nothing: **always state daytime or day + night.**

The bias (nMBE): effect of the two rules

Rule 1 (night): when the Sun is below the horizon ($\theta_z > 90^\circ$), the forecast is 0. **Rule 2:** a negative forecast is replaced by 0.

nMBE: mean bias, as % of mean GHI (see “Measuring the error”). Test 2007.

Model	nMBE no rules, day	nMBE no rules, day + night	nMBE with rules, day	nMBE with rules, day + night
Simple persistence	-0.88 %	0.00 %	-0.88 %	-1.02 %
Clear sky	+16.44 %	+16.28 %	+16.44 %	+16.28 %
Smart persistence	+0.35 %	0.00 %	+1.25 %	+1.11 %
AR(2)	+1.49 %	-0.24 %	+2.02 %	+1.87 %
ELM	+0.69 %	-0.38 %	+1.13 %	+0.99 %

How to read this table

- **No rules, day + night:** both persistences have an nMBE of **exactly zero**.
- **Rule 2:** setting negative forecasts to 0 **raises** the mean (smart persistence: +0.35 % → +1.25 % in daytime).
- **Rule 1:** the nMBE of simple persistence goes from 0 to **-1.02 %** over day + night (next slide).
- **Clear sky:** +16 %: it always forecasts sunshine.

An nMBE depends on how it is computed: **always state** daytime or day + night, with or without rules.

Why does persistence go from 0 to -1 %?

	second-to-last daytime hour	last daytime hour	first night hour
Measurement	150	43	0
Persistence, no rules	...	150	43
With rule 1 (night)	...	150	0

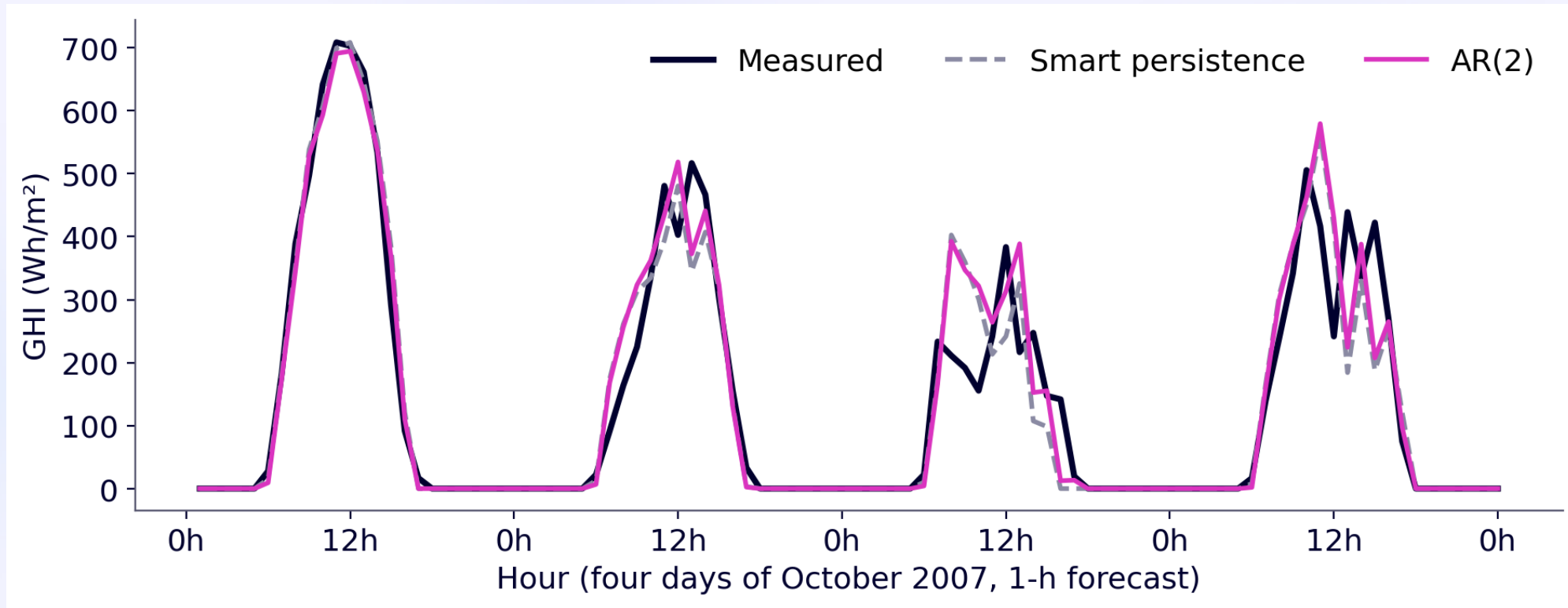
Values in Wh/m²; 43 Wh/m² = 2007 mean of the last daytime hour.

- **No rules:** persistence copies the previous hour's measurement. Over the year, each measurement is used once as measurement and once as forecast: the two sums are equal, **nMBE = 0**.
- **With rule 1 (night):** at the first night hour the Sun is below the horizon, so the forecast 43 is replaced by 0. Every day, a forecast of about 43 Wh/m² disappears.
- Over the year: $365 \times 43 \approx 15,642$ Wh/m², plus 1,008 Wh/m² of small values measured at night: **16,650 Wh/m²** are missing.
- Relative to the GHI measured over the year (1,630,161 Wh/m²): **1.02 %**, hence a day + night nMBE of -1.02 %.

A single hour per day, but it weighs heavily: the day + night mean is only 186 Wh/m², because night hours count as zero.

The rule is physically right (no Sun at night), but it changes the bias.

Forecast and measurement, hour by hour



Ajaccio, October 2007: 1-h forecast.

Both models follow regular days and miss sudden changes: an arriving cloud cannot be seen in the past.

Limits of the AR model

AR(p) assumes a **linear** relation between the CSI and its lags. Often enough at short horizons, but not always.

Difficult situations

- fast clear / overcast transitions (cloud fronts);
- cloud-edge effects ($GHI > GHI^{cs}$);
- very variable sky (cumulus);
- nonlinear links between lags.

Under-fitting

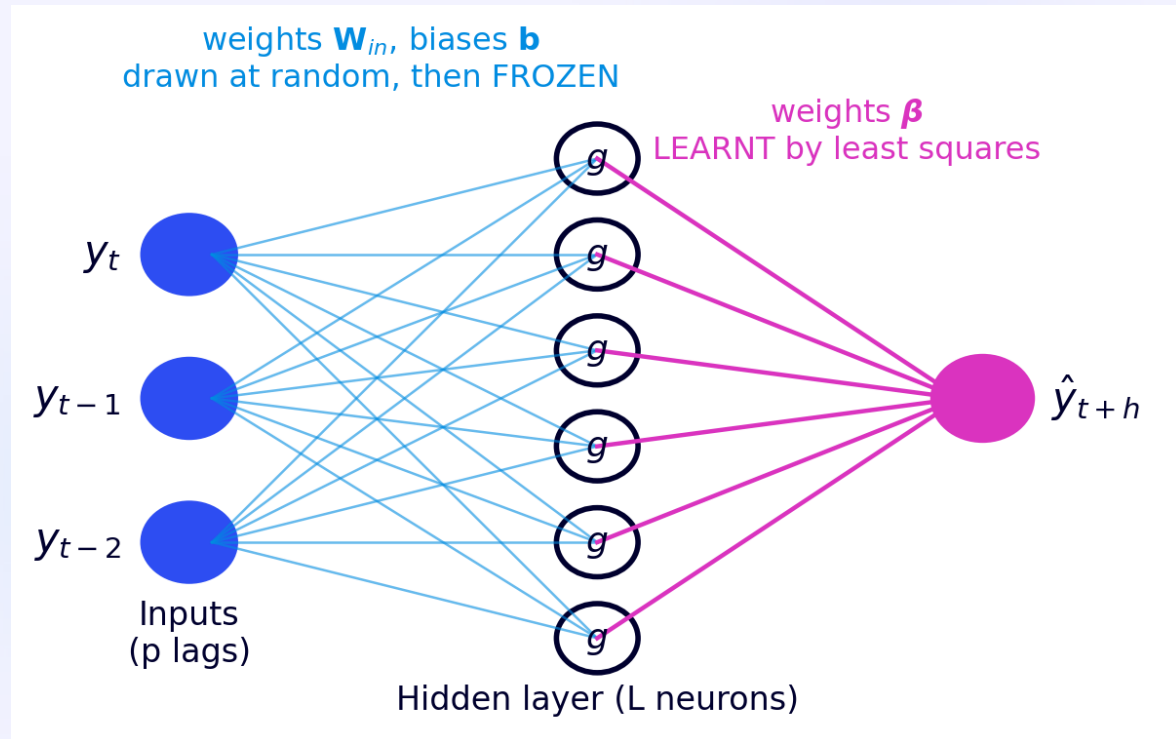
A linear model cannot capture a curved relation, even with many lags.

Saturation

Beyond some p , adding lags no longer improves anything (slide “Choosing p ”).

Idea: transform the inputs nonlinearly, then keep least squares. That is the **ELM**.

Extreme Learning Machine (ELM): the principle



One-hidden-layer ELM (Huang et al., 2006).

- A neural network with **a single hidden layer**.
- Input weights $\mathbf{W}_{in}$ and biases $\mathbf{b}$ are **drawn at random, then frozen**.
- Each neuron applies a nonlinear function g (tanh, sigmoid...).
- Only the output weights β are learnt, **by least squares**.

No gradient back-propagation: training is done **in a single solve**.

ELM: the formulation

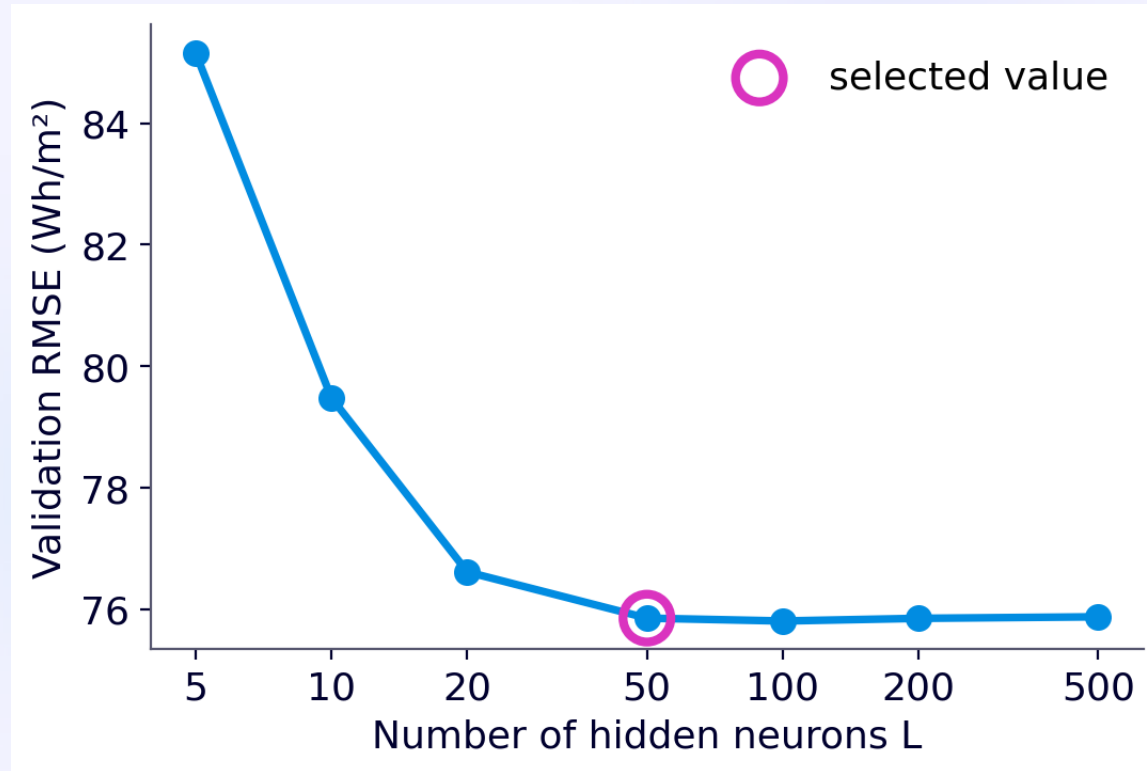
The p lags of each row go through the hidden layer:

$$\mathbf{H} = g(\mathbf{X} \mathbf{W}_{\text{in}} + \mathbf{b}) \quad (\mathbf{N} \times \mathbf{L})$$

	Model	Least-squares solution	Columns used
AR	$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$	$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$	lags of the centred CSI
ELM	$\mathbf{y} = \mathbf{H}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$	$\boldsymbol{\beta} = (\mathbf{H}^T \mathbf{H} + \alpha \mathbf{I})^{-1} \mathbf{H}^T \mathbf{y}$	outputs of the L neurons

Same equation, only the columns change. The small term $\alpha \mathbf{I}$ (*ridge* regression) stabilizes the solve when L is large; at Ajaccio, $\alpha = 10^{-3}$.

ELM: choosing the number of neurons L



Validation RMSE (2006, daytime hours) versus L, for $p = 2$.

- $L = 5$: $85.2 \text{ Wh/m}^2 \rightarrow$ **under-fitting**.
- $L = 50$: 75.9 Wh/m^2 .
- Beyond, no further gain: 75.8 at $L = 200$, 75.9 at $L = 500$ (risk of **over-fitting**).

As for p : test a few values (5 to 500) on validation and keep the smallest near-optimal one. Here **L = 50**.

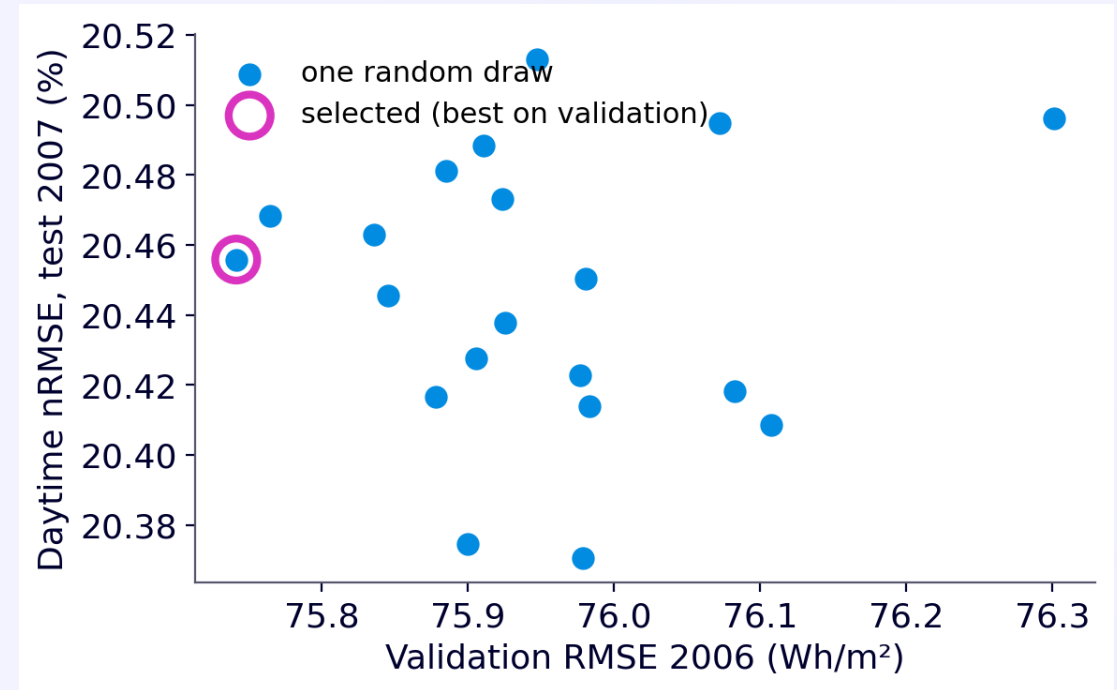
ELM: a result that depends on the draw

- Input weights are random: two trainings give two different models.
- **Candidate search:** make 20 draws and keep the one with the smallest **validation** error (2006).
- The test (2007) is used only at the end, for the selected draw.

Selected draw: test nRMSE **20.46 %**.

The 20 draws: from 20.37 to 20.51 % (mean 20.45 %).

The best on validation is not necessarily the best on test: the gap is small here, but it must be checked.



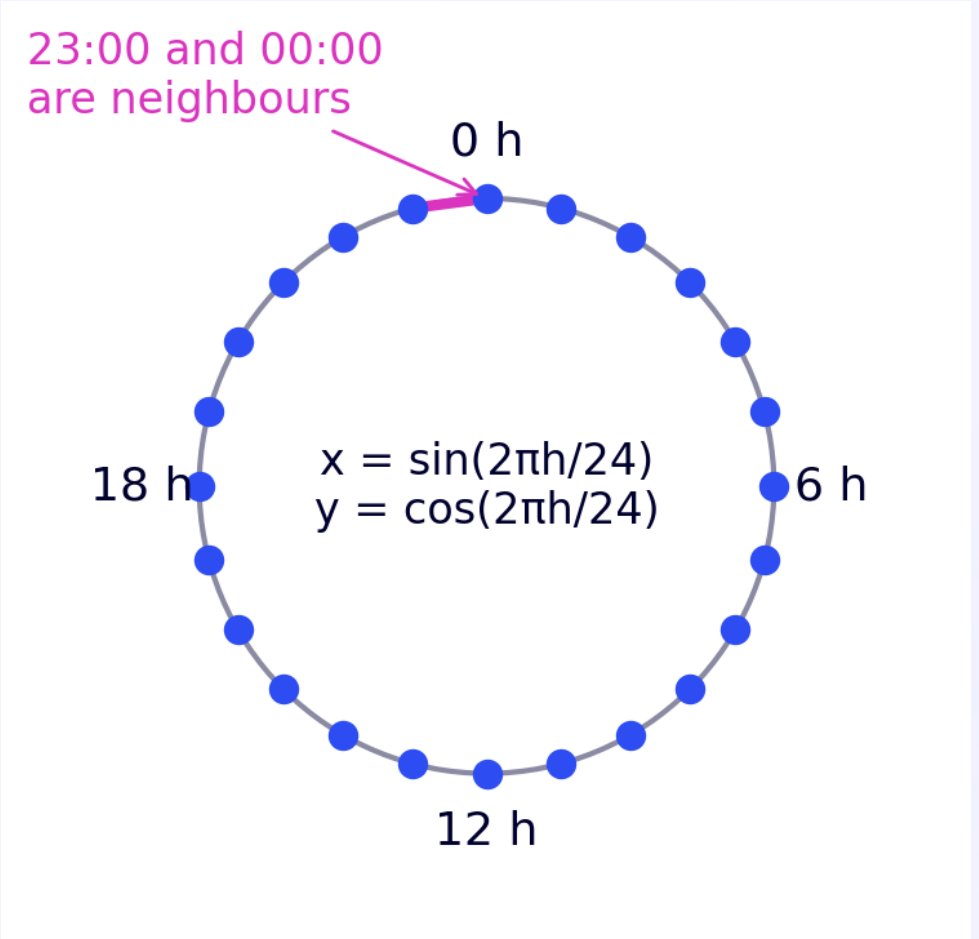
Each point: one draw. Validation 2006 (horizontal), test 2007 (vertical).

Giving the model the time: sin / cos encoding

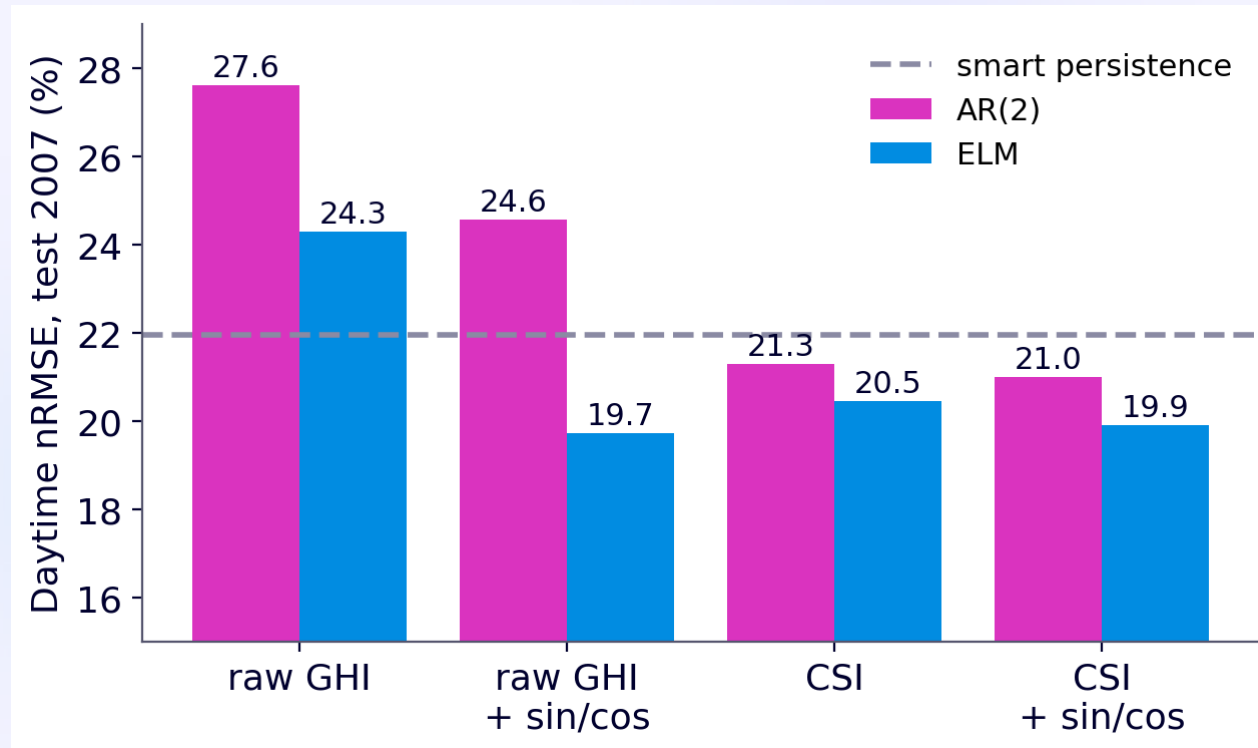
- Idea: add **calendar variables** known in advance to the lags: the hour h and the day of year j .
- Problem: coded as a number, 23:00 is “far” from 00:00, although they are neighbouring hours.
- Solution: place the hour on a **circle**, with two variables:

$$(\sin(2\pi h / 24); \cos(2\pi h / 24)) \text{ and } (\sin(2\pi j / 365); \cos(2\pi j / 365))$$

4 more columns in $\mathbf{X}$. The model then “knows” the hour and the season (Voyant et al., 2025a).



Stationarize or give the time? AR and ELM



Daytime nRMSE, test 2007, same protocol for the 8 variants. “No clear-sky” approach: Voyant et al., 2025a.

Inputs	AR(2) nRMSE	AR NICE ^Σ	ELM nRMSE	ELM NICE ^Σ
raw GHI	27.6 %	0.76	24.3 %	0.66
GHI + sin/cos	24.6 %	0.68	19.7 %	0.54
CSI	21.3 %	0.58	20.5 %	0.55
CSI + sin/cos	21.0 %	0.57	19.9 %	0.54

AR (linear): needs the CSI. On raw GHI, even with sin/cos, it does worse than smart persistence.

ELM (nonlinear): with sin/cos, it learns the Sun's cycle by itself, from raw GHI.

nRMSE and NICE^Σ give the same ranking.

AR or ELM? The verdict

ELM strengths

- fast training, a single solve;
- captures nonlinear relations;
- same method (OLS) as AR.

ELM limits

- L must be chosen;
- result varies with the draw;
- less interpretable than AR;
- over-fits if L grows large;
- needs a lot of data.

At Ajaccio, 1 h ahead

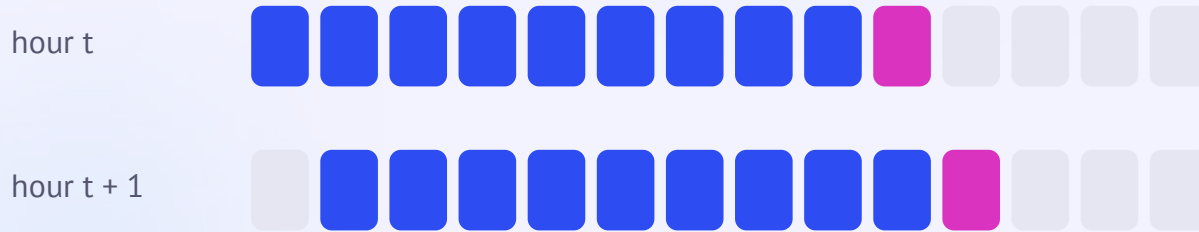
- AR(2): 2 coefficients; nRMSE 21.3 %.
- ELM: 51 learnt + 150 random; nRMSE 20.5 %.
- Skill score SS: +0.03 (AR) versus +0.07 (ELM).

About **100 times more coefficients** (learnt + random) for 4 more points of skill score. Is the gain worth the complexity?

Online learning (1/2): the sliding window



- Trick: no need to recompute everything. $\mathbf{X}^T\mathbf{X} \leftarrow \mathbf{X}^T\mathbf{X} + \mathbf{x}_{\text{new}}\mathbf{x}_{\text{new}}^T - \mathbf{x}_{\text{old}}\mathbf{x}_{\text{old}}^T$.
- Same for the ELM, with $\mathbf{H}$ instead of $\mathbf{X}$.



blue: the W training hours; pink: the hour to forecast

- Matrix $\mathbf{X}$ **always keeps W rows**: the last W hours.
- Every hour, the newest row is **added** and the oldest **removed**, then ridge is re-estimated:

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

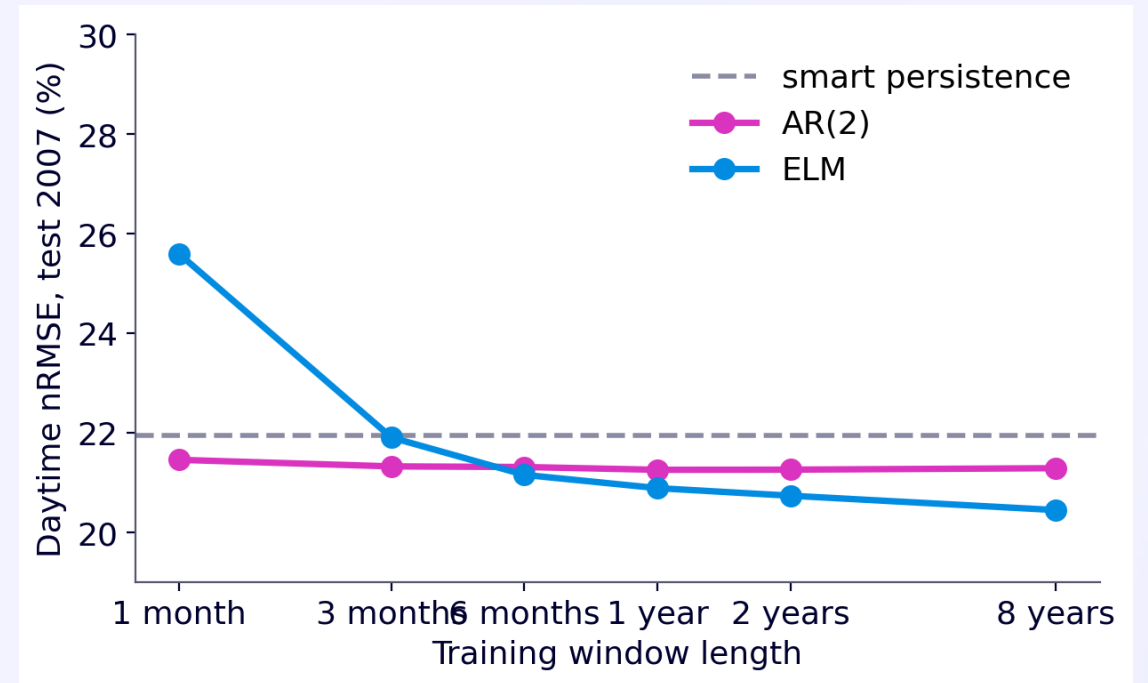
Update: 11 μs for AR, 38 μs for the ELM, per hour.

Useful when the system **changes** (drifting sensor, soiled panel, new site): the model forgets the distant past.

Online learning (2/2): which window length?

- Test 2007: every hour, ridge re-estimated on the last W hours, from 1 month to 2 years.
- Short window: follows recent changes... but learns from little data.

Window W	AR(2)	ELM ($L = 50$)
1 month	21.5 %	25.6 %
3 months	21.3 %	21.9 %
6 months	21.3 %	21.2 %
1 year	21.3 %	20.9 %
2 years	21.3 %	20.7 %
8 years (fixed)	21.3 %	20.5 %



Daytime nRMSE over 2007 versus window length.

AR (2 coefficients) works even with 1 month. The ELM (51 coefficients) **degrades clearly** with a short window: the more parameters a model has, the more data it needs.

Frugality

A **frugal** model reaches good accuracy with little computation, little data and little maintenance.

Model	Learnt coefficients	Training cost	Data needed	nRMSE day (test)
Smart persistence	0	none	1 hour	22.0 %
AR(2)	2	5 ms	1 month is enough	21.3 %
ELM (L = 50)	51 (+ 150 random)	68 ms per draw	several years	20.5 %
AR, 1-month sliding window	2	11 μ s per hour	720 hours	21.5 %

Timings measured on a laptop: only the orders of magnitude matter.

The right question is not “which model is best?” but “**does the gain justify the cost?**” (computation, data, maintenance, monitoring).

Thank you — your turn: the project

Grid and units

Forecasting GHI is forecasting production.

Radiation

Direct, diffuse; Sun position; clouds.

Time series

Components, horizon, AR(1).

Clear sky and CSI

Solis; $CSI = GHI - GHI^{cs}$, centred.

AR and OLS

$\hat{\mathbf{w}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, choice of p .

Evaluation, ELM, online

Day / day + night, sliding window, frugality.

Cyril Voyant — cyril.voyant@minesparis.psl.eu — Mines Paris – PSL, Centre O.I.E.

Worked example on the web: cyrilvoyant.github.io/teaching/solar-forecasting-basics.html

Glossary

Acronym	Meaning
AR(p)	autoregressive model of order p
ACF / PACF	autocorrelation / partial autocorrelation
AIC / BIC	Akaike / Bayesian information criteria
AOD	aerosol optical depth
CSI	clear-sky index
DHI	diffuse horizontal irradiance
DNI	direct normal irradiance
ELM	Extreme Learning Machine
GHI	global horizontal irradiance
MAE	mean absolute error
OLS	ordinary least squares

Acronym	Meaning
NWP	numerical weather prediction
nRMSE	RMSE normalized by mean GHI
RMSE	root mean square error
MBE	mean bias error
sCV	stochastic coefficient of variation
F	forecastability
NICE	score relative to simple persistence
sin / cos	circular encoding of hour and day
SS	skill score
PHS	pumped-storage hydropower

References

- Huang G.-B., Zhu Q.-Y., Siew C.-K. (2006). Extreme learning machine: theory and applications. *Neurocomputing*, 70, 489–501. doi:10.1016/j.neucom.2005.12.126
- Ineichen P. (2008). A broadband simplified version of the Solis clear sky model. *Solar Energy*, 82(8), 758–762. doi:10.1016/j.solener.2008.02.009
- Voyant C. et al. (2017). Machine learning methods for solar radiation forecasting: a review. *Renewable Energy*, 105, 569–582. doi:10.1016/j.renene.2016.12.095
- Voyant C. et al. (2022). Benchmarks for solar radiation time series forecasting. *Renewable Energy*, 191, 747–762. doi:10.1016/j.renene.2022.04.065
- Voyant C. et al. (2026). Stochastic coefficient of variation: assessing the variability and forecastability of solar irradiance. *Renewable Energy*, 256, 123913. doi:10.1016/j.renene.2025.123913
- Voyant C., Despotovic M., Notton G., Saint-Drenan Y.-M., Asloune M., Garcia-Gutierrez L. (2025a). On the importance of clearsky model in short-term solar radiation forecasting. *Solar Energy*, 289, 113490. doi:10.1016/j.solener.2025.113490
- Voyant C. et al. (2025b). NICEk metrics: unified and multidimensional framework for evaluating deterministic solar forecasting accuracy. *Sustainable Energy Technologies and Assessments*, 81, 104588. doi:10.1016/j.seta.2025.104588
- Hyndman R. J., Athanasopoulos G. (2021). *Forecasting: Principles and Practice*, 3rd ed. OTexts. otexts.com/fpp3
- Box G. E. P., Jenkins G. M., Reinsel G. C., Ljung G. M. (2015). *Time Series Analysis: Forecasting and Control*, 5th ed. Wiley.
- Data: Météo-France, Ajaccio station. All figures and schematics drawn by the author.

Appendix A — Wide-sense stationarity

A process (z_t) is **wide-sense stationary** if:

$$E[z_t] = \mu \text{ (constant)} \quad \text{and} \quad \text{Cov}(z_t, z_{t+k}) = \gamma(k) \text{ (depends only on the lag } k)$$

In practice

- the mean is constant over time;
- the variance is finite and constant;
- the dependence between two instants depends only on their gap.

Simple check: plot the mean and standard deviation month by month (slide “Is the CSI stationary?”). Statistical tests exist but are not needed here.

Appendix B — The best possible forecast

$$\hat{y}_{t+h} = E[y_{t+h} \mid y_t, y_{t-1}, \dots, y_{t-p+1}]$$

- Among all forecasts built from the past, the one that minimizes the mean squared error is the **conditional expectation**: the mean of possible futures, given the past.
- We then write $y_{t+h} = \hat{y}_{t+h} + \varepsilon_{t+h}$: predictable part + error, with a zero-mean error.
- An AR(p) model is a **linear** approximation of this expectation; the ELM, a **nonlinear** one.

If the model errors form white noise (non-significant error ACF), no usable information is left in the past.

Appendix C — Choosing p with AIC or BIC

$$\text{AIC} \approx N \cdot \ln(\hat{\sigma}^2) + 2(p+1)$$

$$\text{BIC} \approx N \cdot \ln(\hat{\sigma}^2) + (p+1) \cdot \ln N$$

- $\hat{\sigma}^2$: variance of the training errors;
- keep the p that gives the smallest criterion.

- The first term falls as the model fits better; the second **penalizes** each added coefficient.
- BIC penalizes more than AIC: it picks smaller models.
- Advantage: no validation period needed. Drawback: assumes Gaussian errors.