Continous Causal States and some other ideas

Nicolas Brodu

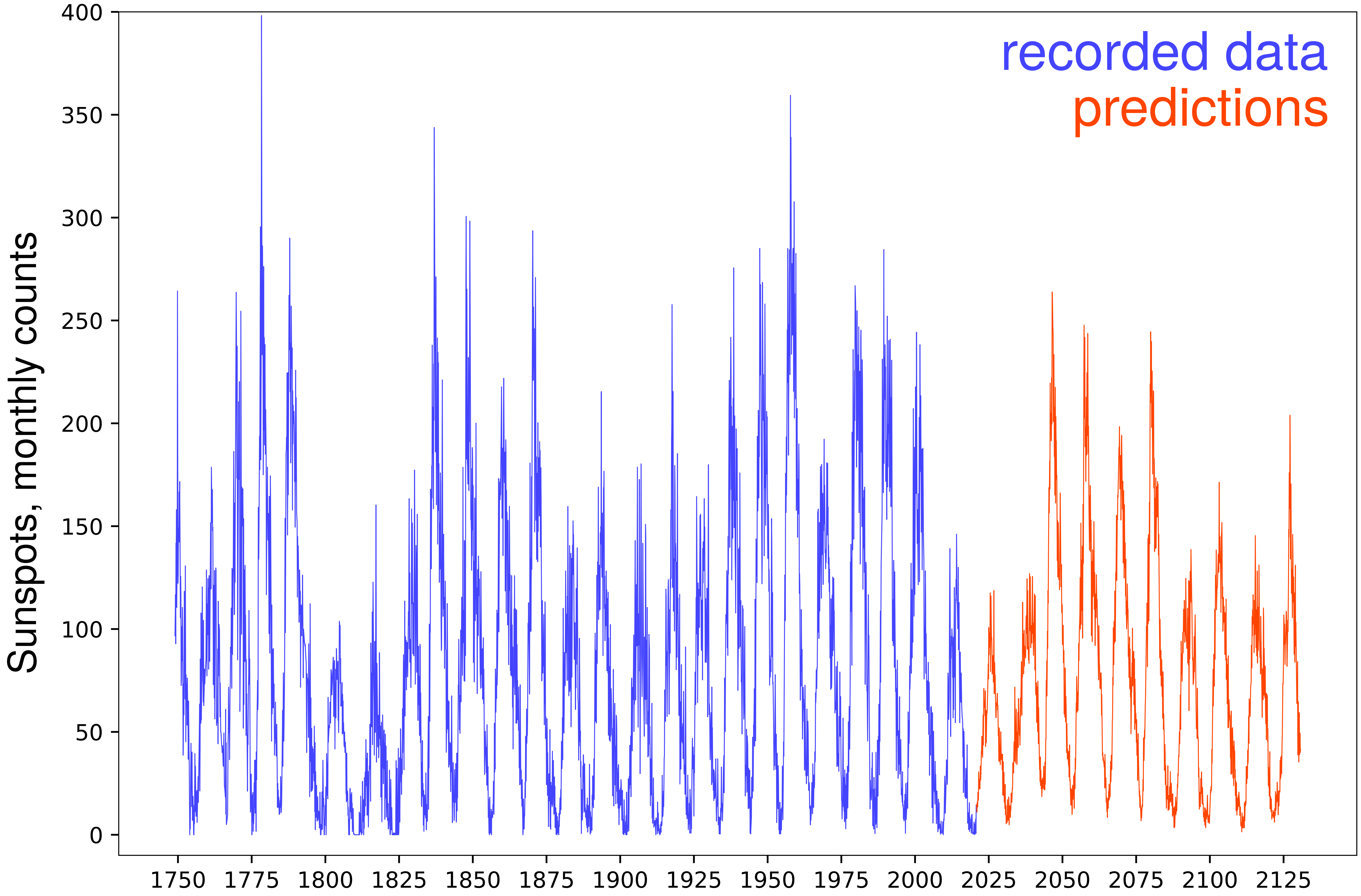
nicolas.brodu@inria.fr

Inria Bordeaux, France

Inference for dynamical systems meeting
Feb 2021

(slightly updated after the presentation)

Motivating results : Sunspots time series



Known to be extremely hard to predict

(these predictions should not be trusted!)

We can try anyway

Illustrative example for the method

Motivating results : Sunspots time series (notations)

Notations (slight change from Adam's)

– $v_t \in \mathcal{V}$: the observed values at each time

– $x_t \in \mathcal{X}$: the observed past time *series* : $x_t = v_{\tau \leq t}$, possibly truncated : $\tau > t - L^X$

– $y_t \in \mathcal{Y}$: the observed future time *series* : $y_t = v_{\tau > t}$, possibly truncated : $\tau \leq t + L^Y$

– X, Y : Random variables for the time series

For the Sunspots example

– Observable: Sunspot counts

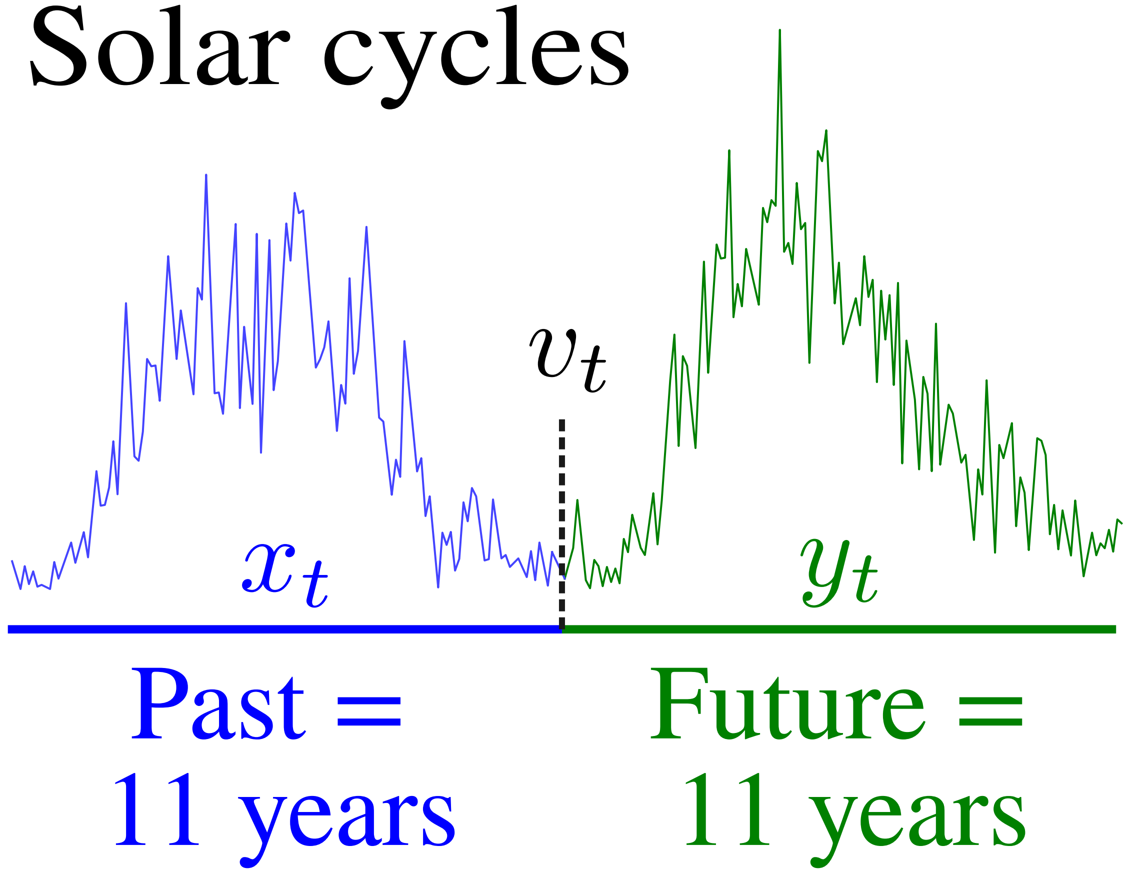
– Measurements: Monthly total… during day time.
Averaged over multiple observatories.

– Discrete time

– Continuous values (due to averaging) $v_t \in \mathcal{V}$

– Temporal scale for X, Y : 1 solar cycle

Solar cycles



$L^X = L^Y = 132$ months

Different views on dynamical systems – Causal States

Basic dynamical systems view

$v_t = V(\omega_t)$ observable value (sunspots)

ω_t unobservable system (full Sun) state

$v_t = (U^t V)(\omega_0)$ Koopman operator

$v_{\tau > t}$ values to predict. Can use U^τ

Markov order- L^X process view

$x_t = X(\omega_t)$ observable value (series)

$x_t = (U^t X)(\omega_0)$ Koopman operator

predictions on future *x* values

⇒ focus remains on the values

Note: *x* can also be seen as a vector of time-lagged values with lag=1

Causal states focus on distributions $P(Y|X)$

X should include all the past that has some (causal) effect on the present $L^X \to \infty$ or not

Y should include all the future that is influenced by the present $L^Y \to \infty$ or not

$s_t \equiv P(Y|X = x_t)$ distribution of possible futures. Defines a partition of $\mathcal{X}$

No new observation can distinguish x' from x in the same causal state → Markovian as consequence

$s_t = (U^t S)(\omega_0)$ evolution operator ?

$\mathbb{E}_P[f(y)]$ Expectation operator makes predictions. f could be X, or any quantity of interest

} detailed shortly

Motivating results : Sunspots time series (attractor)

Reconstructed dynamics / attractor ⟶

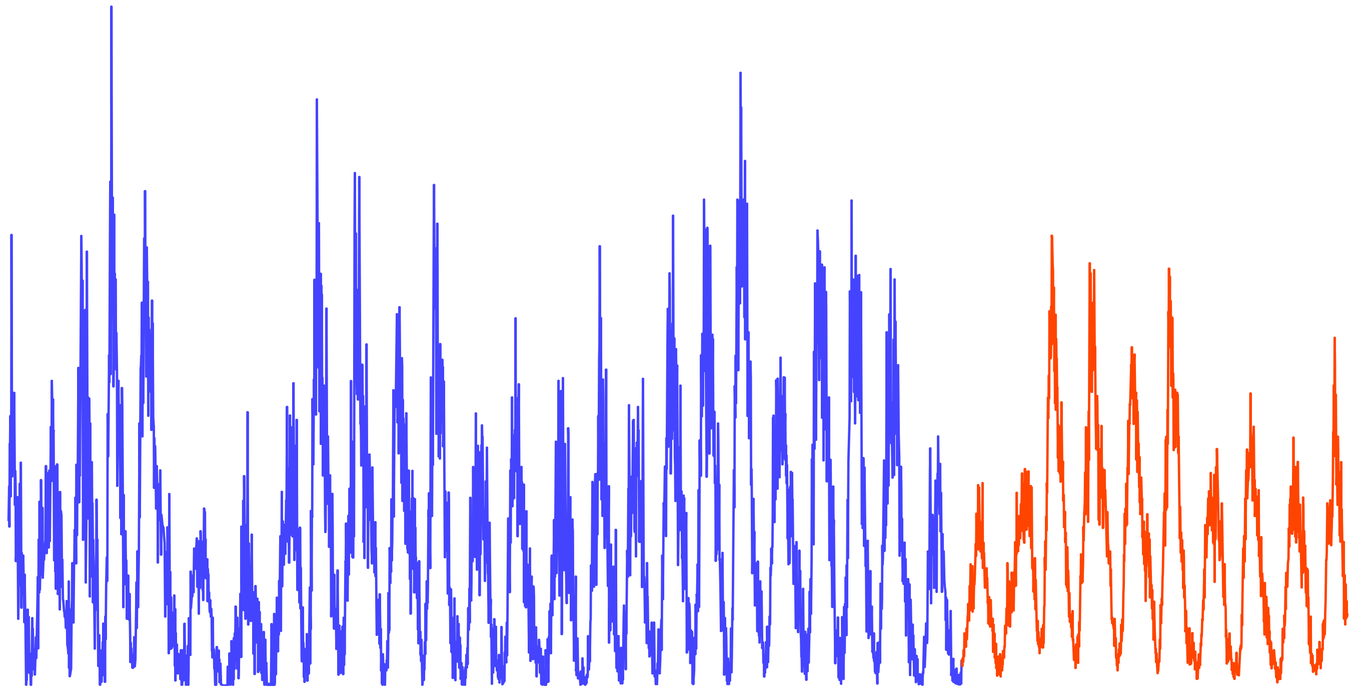
– Each point ⇔ causal state

– “Projection” from space of distributions

Axes = most relevant variables

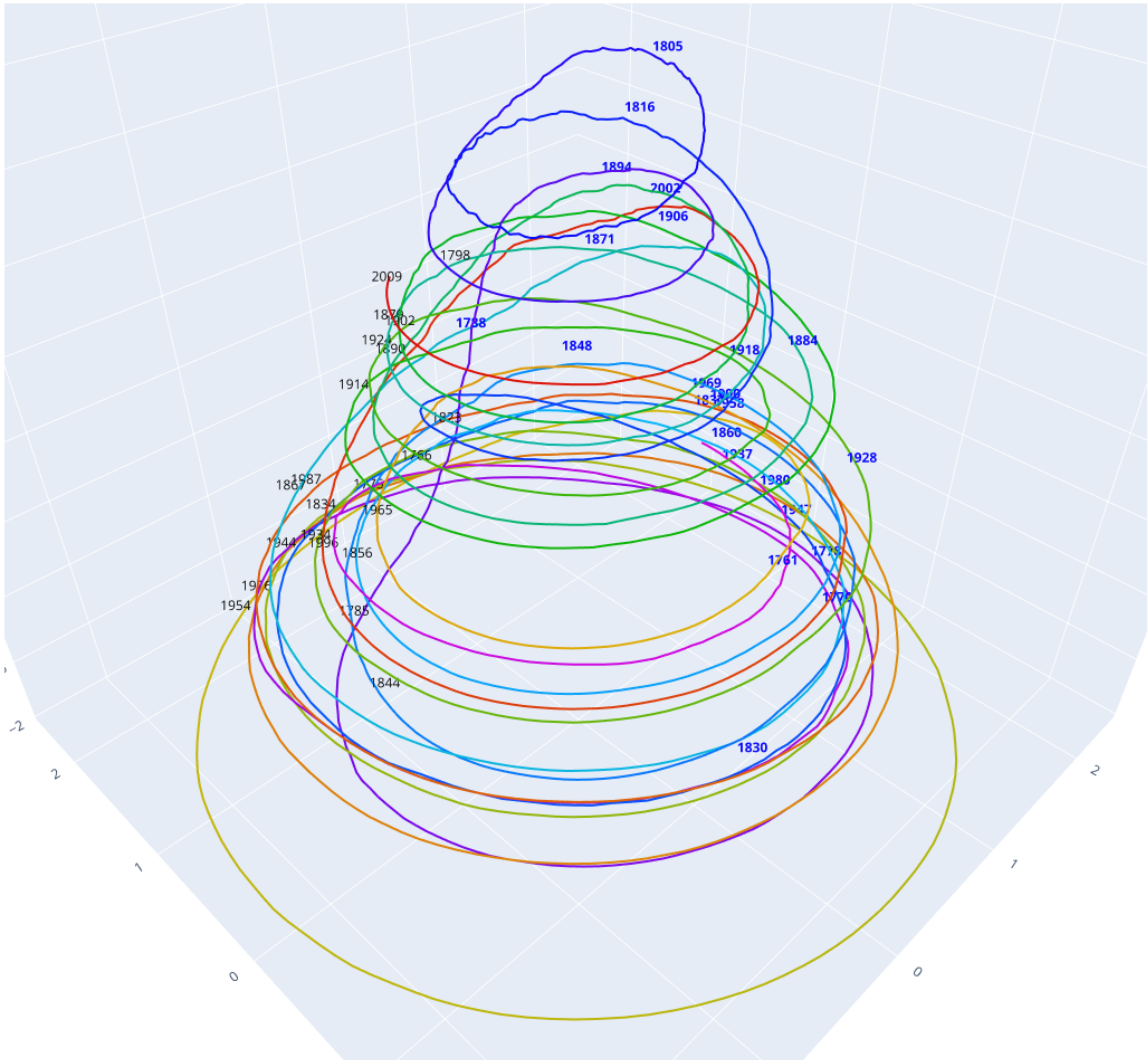
– axis 1 & 2 : 11-years cycle and phase

– axis 3 : amplitude modulations over
80-100 years (= Gleissberg cycles)

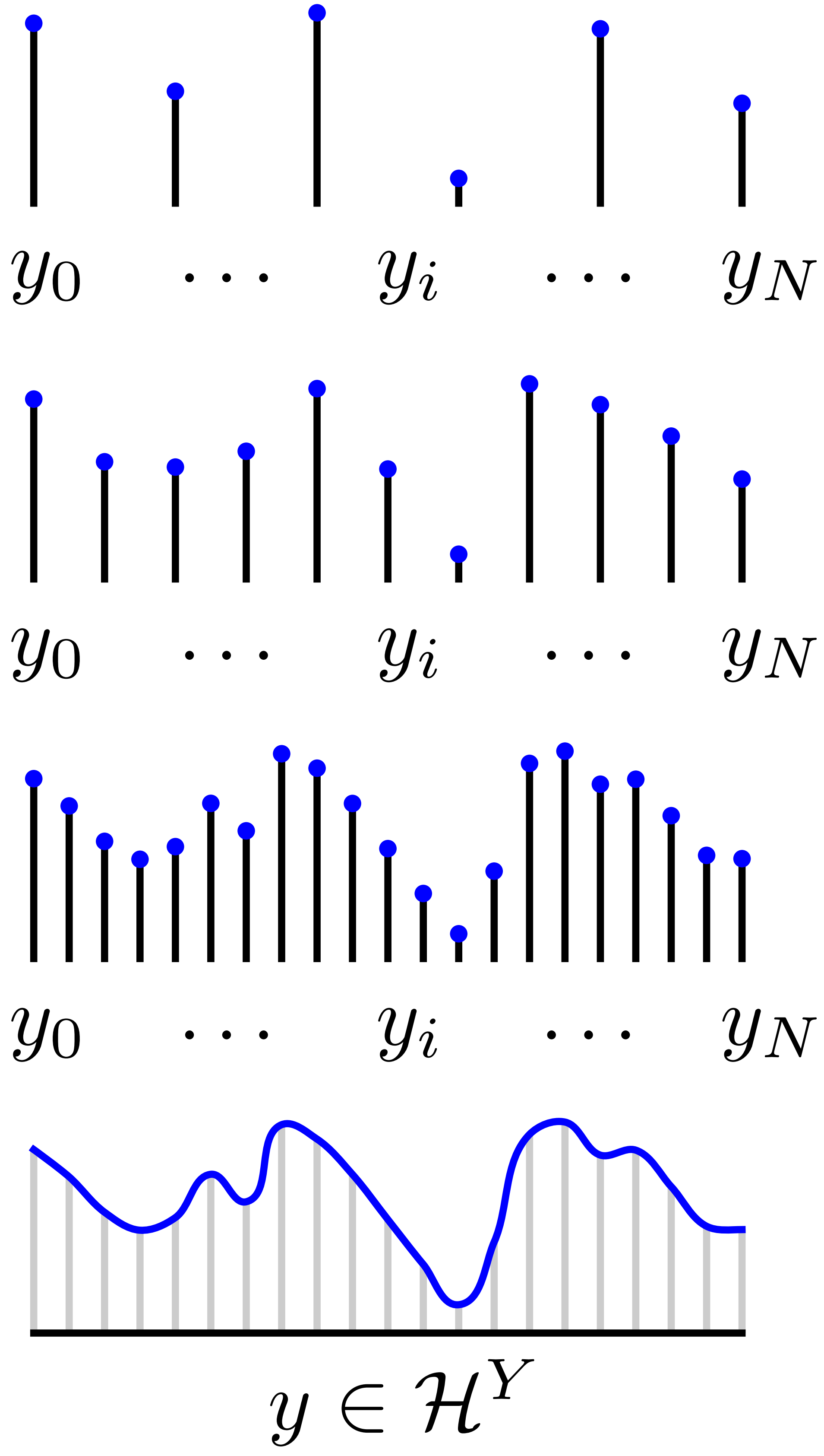


Pattern found ≫ analysis scale

⇒ Evolution operator encodes the dynamics
(and uses it for predictions)



Functions defined by their values, seen as ∞-dimension vectors



N=5 Vector of dimension 5
= Class of functions with these values

Note: operators acting on causal states become matrices

N=10

Increasing number of points
=
Increasing the dimension
=
Finer and finer equivalence classes of functions

Causal state as a vector defined on data samples

Coarser numerical approximations

N=20

N→∞ Single function, but now lives in an ∞-dimension space
Causal state = distribution of Y = function of the Y

In terms of Kolmogorov complexity

Program for generating a string ⇔ Formal expression for function values
Most strings are random ⇔ Most functions of data have no expression

Reproducing kernels, distributions

Analogy with L^2

– Inner product $\langle f, g\rangle_{L^2} = \int_Y f g \,\mathrm{d}\mu^Y$

– Delta selects an element $\langle f, \delta_y\rangle_{L^2} = f(y)$

– Delta as a function of 2 variables $\delta_y = \delta(y, \cdot)$

– $\delta(y_1, y_2) \neq 0$ indicates equality

Reproducing kernel in Hilbert Space $\mathcal{H}^Y$

– Inner product implicitly defined $\langle f, g\rangle_{\mathcal{H}^Y}$

– Kernel "reproduces" an element $\langle f, k_y\rangle_{\mathcal{H}^Y} = f(y)$

– Kernel as a function of 2 variables $k_y = k(y, \cdot)$

– $k(y_1, y_2)$ indicates the similarity between y_1, y_2

Kernels act as generalized δ : yes/no equality → similarity

Any positive symmetric definite function is a kernel for an associated Hilbert Space

Widely used example $k(y_1, y_2) = \exp\left(-\|y_1 - y_2\|^2\right)$

Aronszajn 1950

Representing distributions

– Use the data span as a pseudo-basis : $P(Y) \widehat{=} \sum_{i=1}^{N} c_i\, k(y_i, \cdot)$

Distribution estimated as a vector c of N elements

– Unconditional distributions : $c_i = 1/N$ → usual kernel density estimation

Gretton *et al* 2012

– Conditional distributions $P(Y|X = x)$: c_i depends on how x $\underbrace{\text{is similar to}}_{\text{involves } k^X(x, x_i)}$ observation x_i

Causal states – kernelized version

Causal states = distribution on *series*

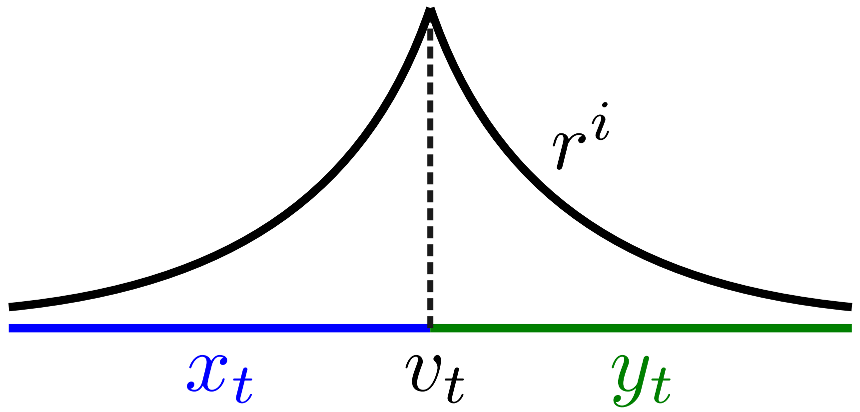
$P(Y|X=x) \widehat{=} \sum_{i=1}^{N} c_i \, k^Y(y_i, \cdot)$

$x_t = (v_\tau)_{-L^X < \tau \le t}$ Need a kernel k^X on past series for the definition of c_i

$y_t = (v_\tau)_{t < \tau \le t+L^Y}$ Need a kernel k^Y on future series for the data span in $\mathcal{H}^Y$

Product kernels = kernel of product spaces

Aronszajn 1950



$k^Y(y, y') = \prod_{i=1 \dots L^Y} k_i^V(y_i, y_i')$ with $y_i = v_{t+i\tau}$ the series i-th entry

$k_i^V(y_i, y_i') = k^V(v, v')^{r^i}$ with $k^V(v, v')$ a kernel on values, r a decay ratio for causal influence

Also works for composing heterogenous data sources

E.g., T = temperature, P = precipitations, E = evapotranspiration

$$k^V(v, v') = k^T(t, t') \; k^P(p, p') \; k^E(e, e')$$

An analyzing scale is needed for each data source

$k^V\left(\frac{v}{\lambda}, \frac{v'}{\lambda}\right) = \exp\left(-\left\|\frac{v}{\lambda} - \frac{v'}{\lambda}\right\|^2\right)$ Kernel acts on dimensionless data

Main parameters = scales

For each data source:

– Past causal duration L^X
– Future causal duration L^Y
– Data scale

The nature of the kernel is surprisingly not as important

Another example : Forest ecosystem

Heterogenous measurements

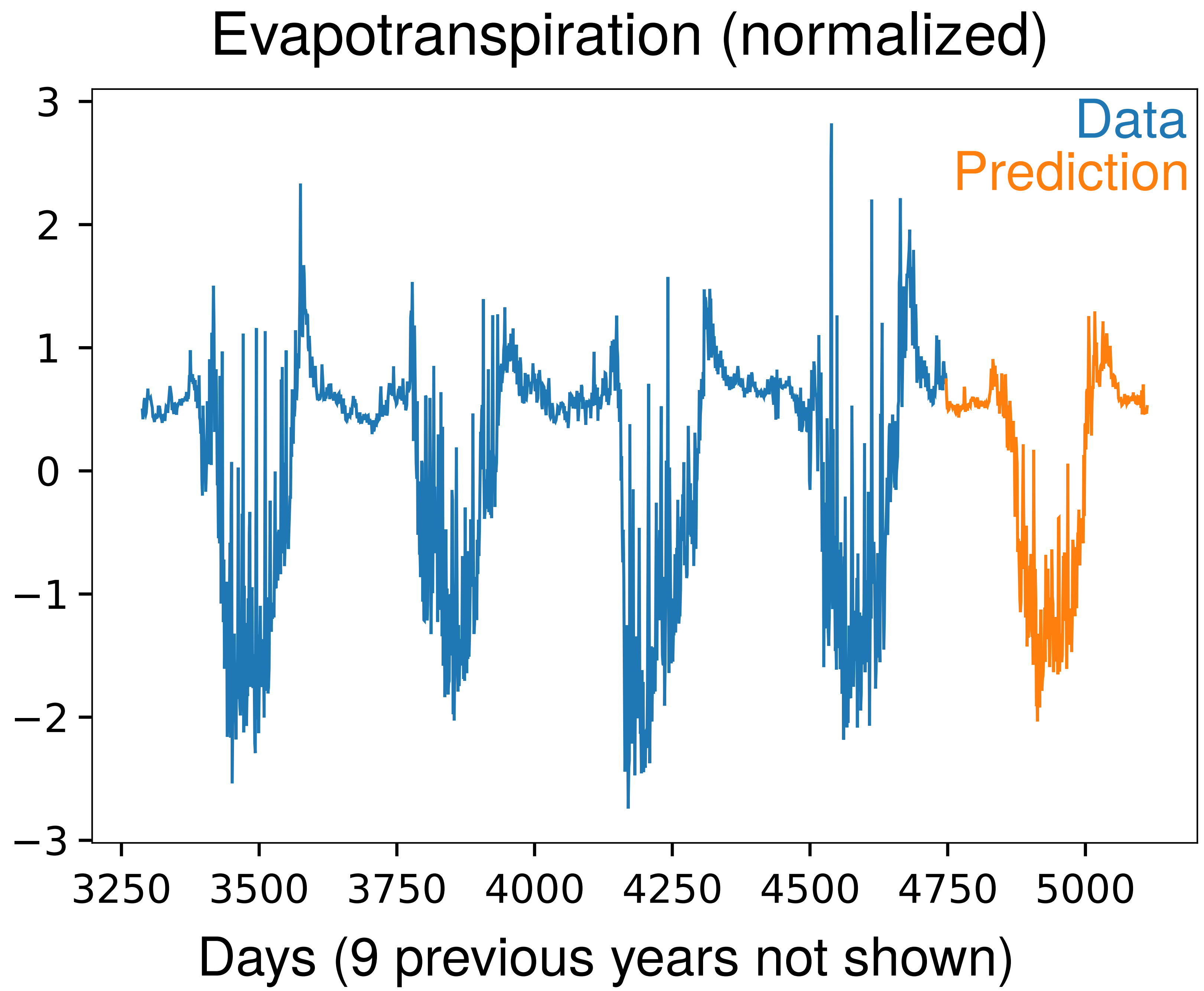
- Temperature
- Solar energy influx
- Precipitations
- Soil water content
- Evapotranspiration
- CO_2 flux

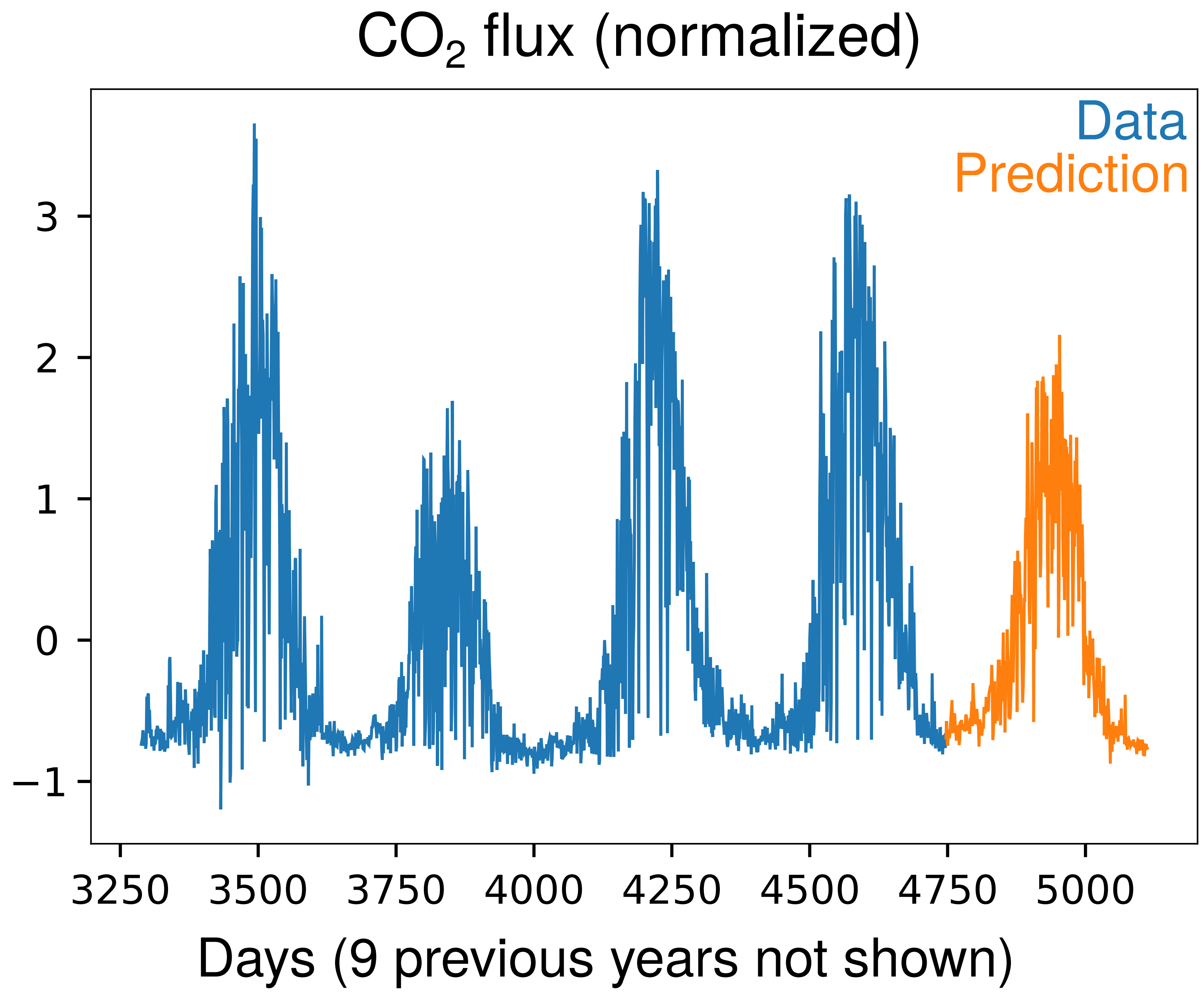
Scales

- Past = 2 weeks
- Future = 1 week
- Data = 10 std.dev.

(need to fix this)

Seasonal patterns $\gg$ analysis scale are clearly captured and predicted





Another example : Forest ecosystem (attractor)

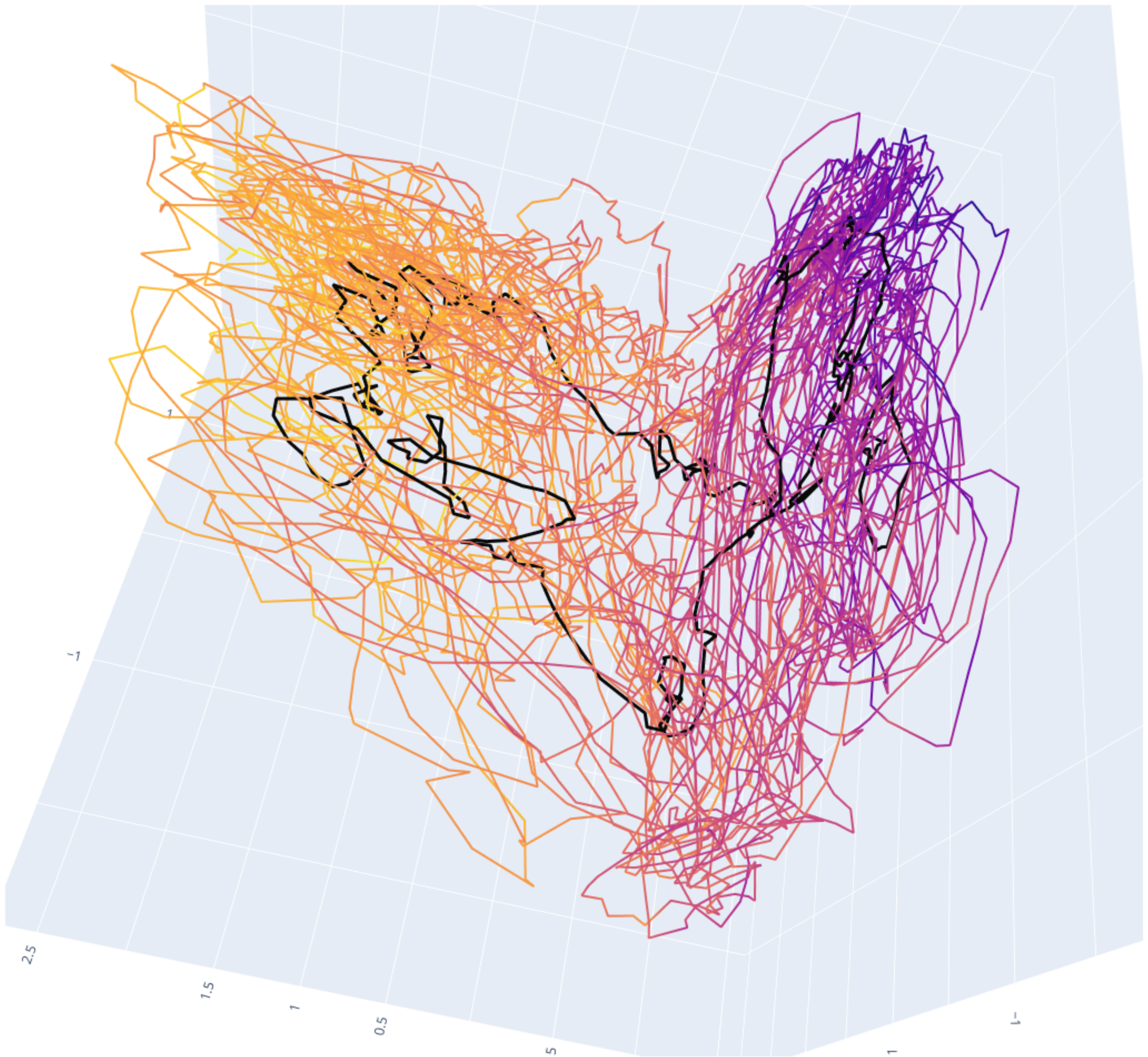
Reconstructed dynamics / attractor ⟶

– Each point ⇔ causal state

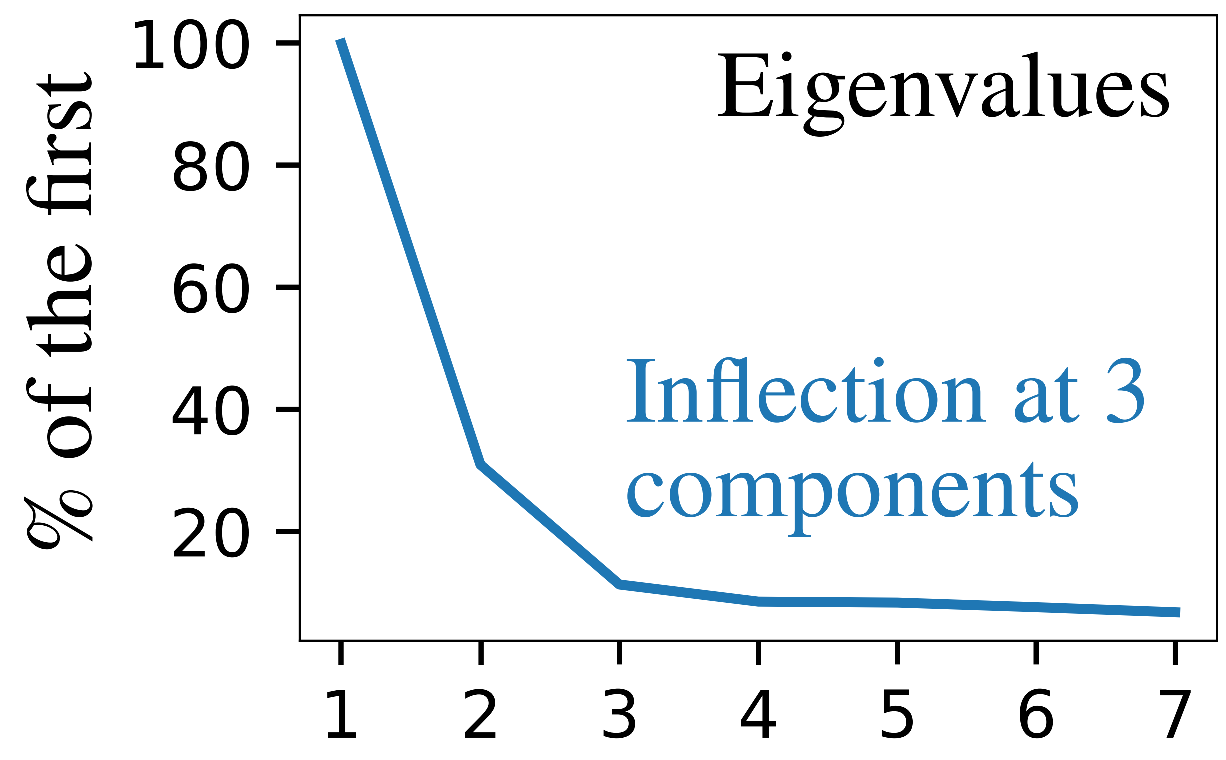
– "Projection" from space of distributions

– Black curve = predicted states

how to do this = next slides!



Number of relevant components ?



Interpretation

– Color = temperature

– Recovers the seasonal cycle

Causal states embedding

Causal states = distribution = point in ∞-dimensional RKHS

$s \equiv P(Y|X=x) \mathrel{\widehat{=}} \sum_{i=1}^{N} c_i \, k^Y(y_i,\cdot)$

$s \in \mathcal{S} \subset \mathcal{H}^Y$ subset is indexed by $x \in \mathcal{X}$

Reproducing property
$k(y,z) = \langle k_y, k_z \rangle$

Geometry of $\mathcal{S}$, the set of causal states

Distances $\|s - s'\|^2_{\mathcal{H}^Y} = \langle s - s', s - s' \rangle$ can be written as a function of $c, c', k^Y(y_i, y_j)$

⇒ Distances between every pair of states can be computed from data! ← thus, the N-1 simplex

⇒ An embedding can be found $\mathbb{S} \subset \mathbb{R}^{N-1} \xleftrightarrow{\text{One to one embedding}} \mathcal{S} \subset \mathcal{H}^Y$

Diffusion Maps recover the geometry independently from the sampling density ← other choices are possible

Low dimension hypothesis

Causal states are intrinsic properties of the physical process (and invariant by coordinate transforms)

⇒ Main structure with M ≪ N descriptive parameters (independent of observation count)
+ small fluctuations / errors (that depend on N)

Diffusion Maps is a spectral method, eigenvalues = how relevant is each dimension (similar to PCA)

Dynamics, inference

Back to basic dynamical system in $\mathbb{S} \subset \mathbb{R}^{N-1}$?

Sensitivity to initial conditions: maybe keep M’ > M components

Yes !

$s \mathrel{\widehat{=}} (\psi_1, \ldots \psi_M \ldots \psi_{N-1})$ is a one-to-one mapping

Wait! Pr on states? Are we going to use a secondary RKHS? No! Push-forward measure from $\mathcal{X}$ OK

$s_{t+1} = U s_t$ with Koopman operator estimation methods

$Q_t = \Pr(s_t)$ and $Q_{t+1} = F Q_t$ with Perron-Frobenius

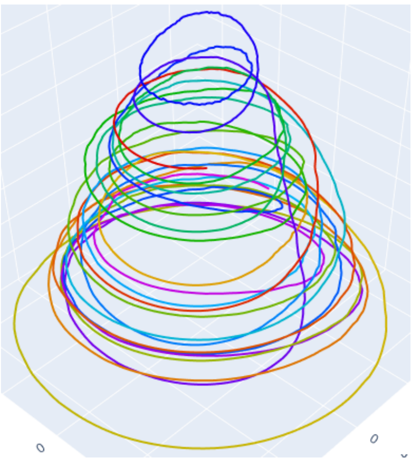
Work in progress… not implemented yet

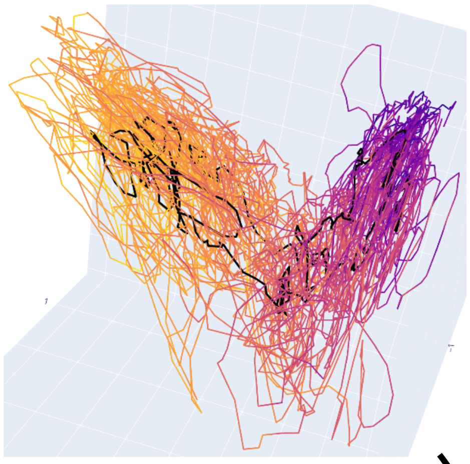
No !

$\mathcal{S} \equiv \mathbb{S}$ is indexed by $\mathcal{X}$: need to guarantee that $U s_t$ remains in $\mathbb{S}$

In particular, $\mathbb{S}$ is not convex $\Rightarrow$ cannot just estimate U, F, with arithmetic averages

The mapping depends on N

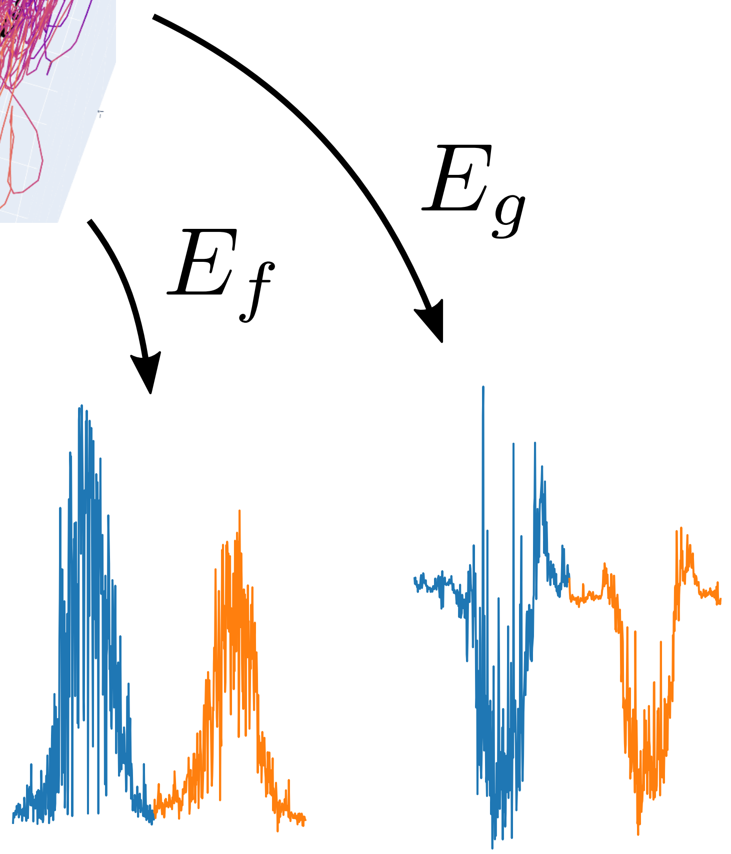




Inference

$s_t \equiv P(Y | X = x_t)$ is a distribution of futures given an observed past.

$E_f = \mathbb{E}_{Q,t} \mathbb{E}_P [f(y)]$ makes predictions for future quantities of interest from the current state (or distribution of states)



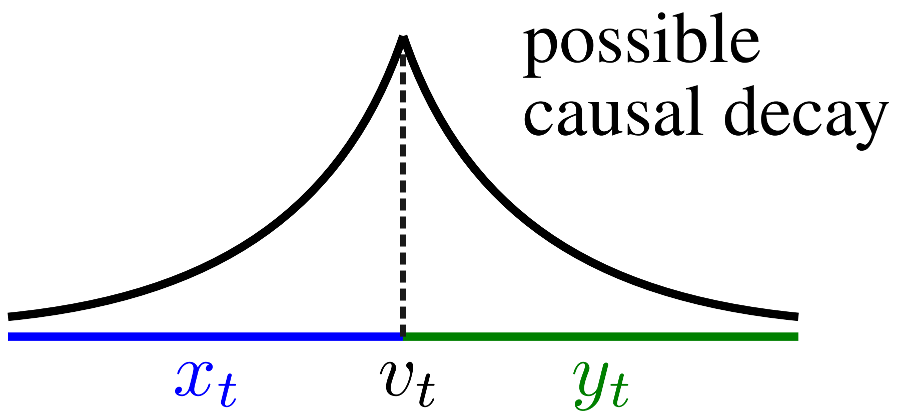
Predictions for any *f* use histories from *all* dependent variables

Part 2

Continous Time
and
Information-related issues

From discrete to continuous time (information perspective)

Definition of causal states



X should include all the past that has some (causal) effect on the present

Y should include all the future that is influenced by the present

$s_t \equiv P\left(Y|X = x_t\right)$ distribution of possible futures

No new observation can distinguish x' from x in the same causal state

Information perspective

Discrete time case: $s_t \to s_{t+1}$ transitions correspond to new information

Discrete data: New symbol

Edge-labeled unifilar transition graph, the ε-machine M^τ transitions $t \to t+\tau$

Continuous data: Motion in the causal state space $\mathcal{S}$

Evolution operators encode the process dynamics U^τ transitions $t \to t+\tau$

Continuous time case: $s_t \to s_{t+\mathrm{d}t}$ transitions correspond to a rate of new information

If that rate is limited: $D_{KL}\left(s_{t+\mathrm{d}t} \| s_t\right) \to 0$ and this implies $\left\|s_{t+\mathrm{d}t} - s_t\right\|_{\mathcal{H}^Y} \to 0$

⇒ Continuous trajectories !

Otherwise, sudden introduction of new information ⇒ jumps

From discrete to continuous time (modeling perspective)

Possible sources of discontinuities (= ∞ information rate)

Fundamental law = information comes in discrete packets — Quantum world

Data is measured at scale ≫ continuum — Renewal process modeling a queue

L^X, L^Y too short ⇒ introduce information jumps — Long range correlation

Continuous-time, continuous state model

Canonical Wiener process for continuous trajectories → $\mathrm{d}W$

Model becomes an inhomogenous Itô diffusion $\mathrm{d}s = a(s)\mathrm{d}t + b(s)\mathrm{d}W$

Evolution of distributions $Q(s)$ using the Fokker-Planck operator $Q(s, t+\tau) = F^\tau Q(s,t)$

$F^\tau = e^{\tau\Gamma}$
with Γ = adjoint of the process generator

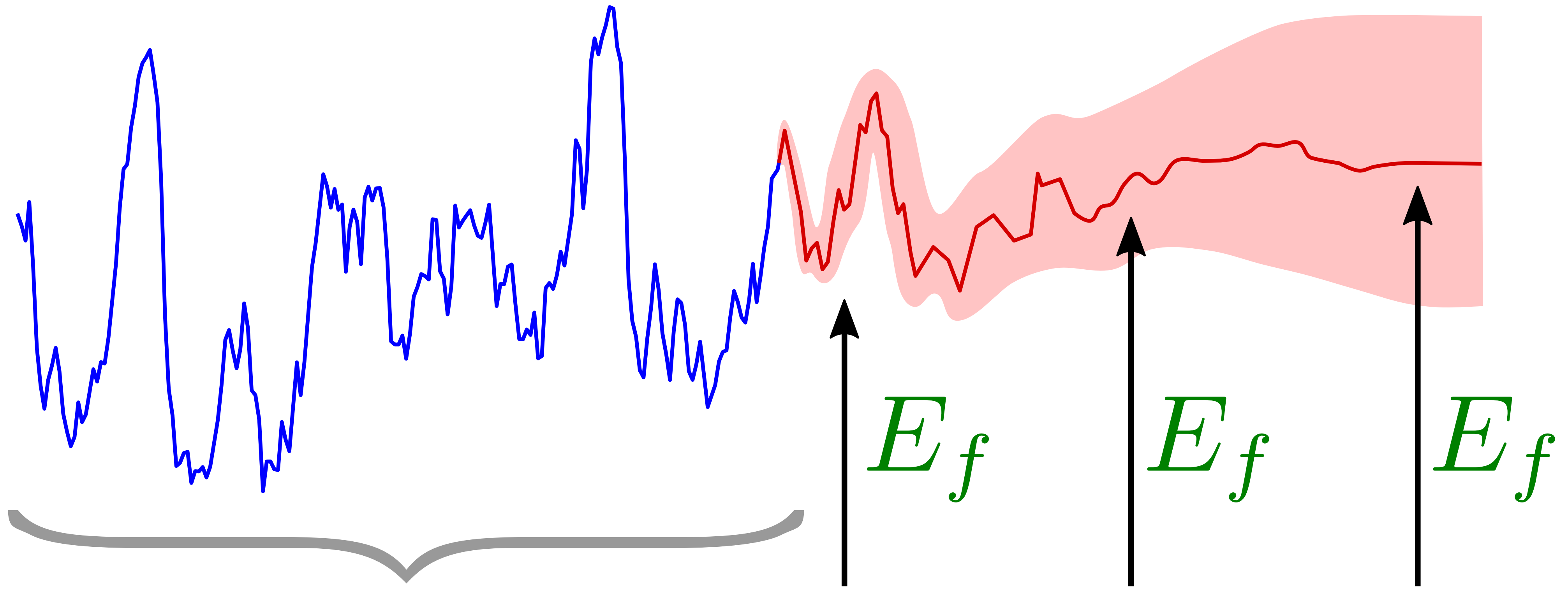
Modeling discontinuities

With a stochastic jump component $\mathrm{d}s = a(s)\mathrm{d}t + b(s)\mathrm{d}W + \mathrm{d}J(s)$

With a Lévy flights, with forced deterministic jump states (as in renewal processes)...

Diffusion of information, loss of prediction accuracy

Current data → **Predictions**



Predictions end up with the long-term average distribution

$$Q_0 \xrightarrow{F_\tau} Q_\tau \dashrightarrow Q_{\tau'} \dashrightarrow Q_\infty$$

Initial information is completely lost

Initial causal state (or distribution of)

Diffusion, distribution of causal states

Limit distribution of the Itô diffusion

This model specifies *how* useful information for prediction is diffused / lost through time

– Average rate of info loss ?

– Information "Half-life" = time scale for accuracy / 2 ?

⇒ To answer with meromorphic calculus and spectral decomposition of F_τ ?

Anomaly detection - quantifying information in states ?

Example: El-Niño anomalies

– 4 sea surface temperature indicators

– Precipitations in 9 regions along the south pacific coast

– Past scale = 2 years

– Future scale = 1 year

– Data scale = 10 standard deviations

(also need to fix this)

Results

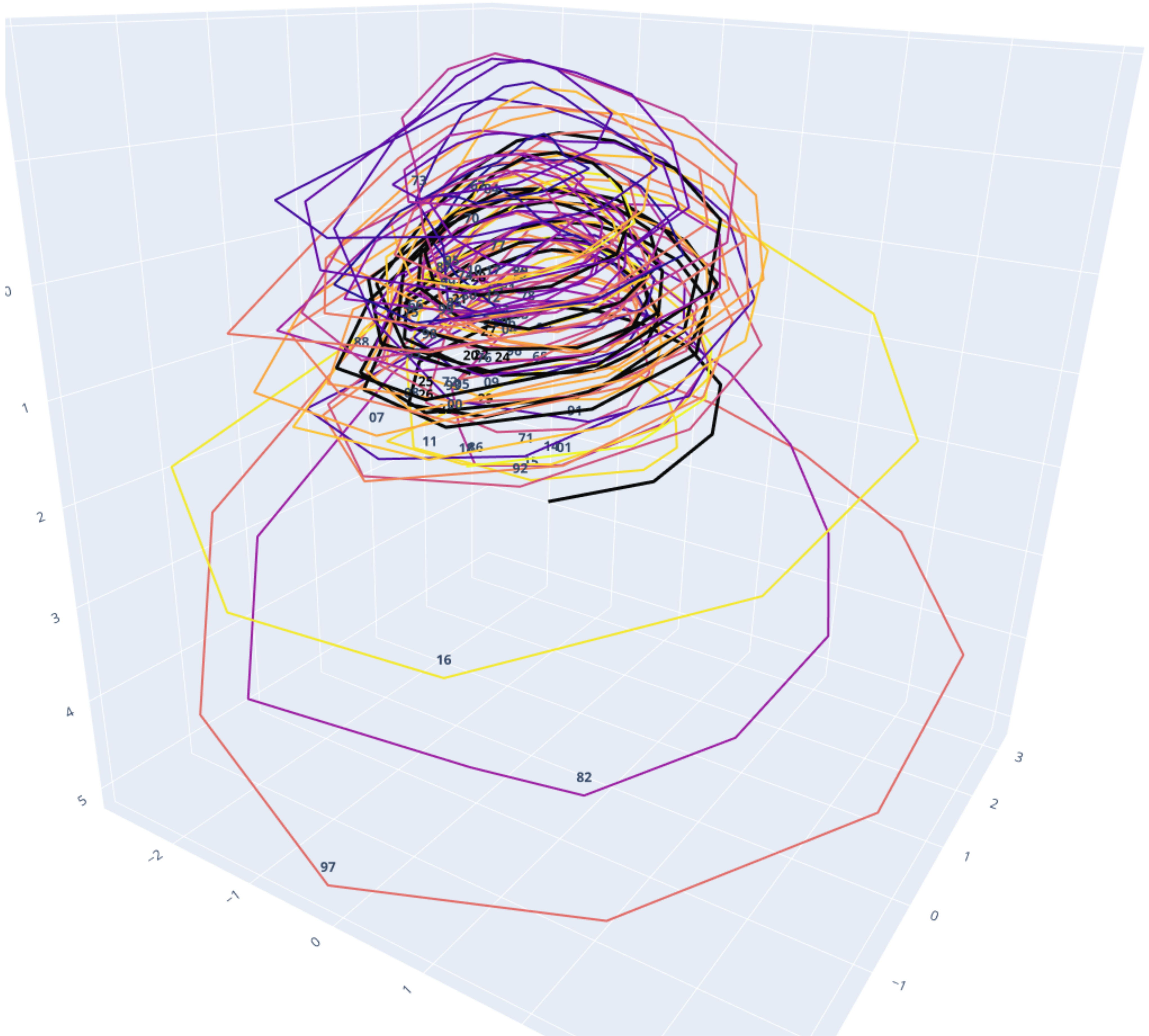
– Seasonal cycle well recovered

– 1982, 1997 and 2016 large events stand out

How to quantify / detect anomalies?

– Automatically (esp. in dim > 3)

– At what scales ? : limit of self-information of causal states → 0 at large scale and → ∞ at small scales



Open project: information spectrum

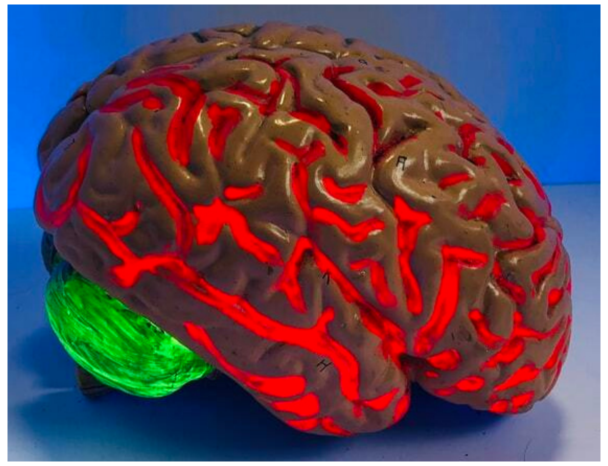
Entropy reduction
needs energy

Information / structure rather than energy dissipation

Energy dissipation allows to maintain patterns (out-of-equilibrium open & dissipative systems)

These patterns often have a *functional* role (e.g. living systems)

$\Rightarrow$ Can we create an "information spectrum", instead of a "power spectrum" ?



May have the same power spectrum,
may dissipate both $\approx$ 30 W,
but their information spectrum should differ