

ENVIRONMENTAL RESEARCH  
LETTERS

## LETTER

## OPEN ACCESS

## RECEIVED

3 January 2022

## REVISED

13 April 2022

## ACCEPTED FOR PUBLICATION

14 April 2022

## PUBLISHED

4 May 2022

Original Content from  
this work may be used  
under the terms of the  
[Creative Commons  
Attribution 4.0 licence](#).

Any further distribution  
of this work must  
maintain attribution to  
the author(s) and the title  
of the work, journal  
citation and DOI.

Physics-aware nonparametric regression models for Earth  
data analysis

Jordi Cortés-Andrés<sup>1,\*</sup> , Gustau Camps-Valls<sup>1</sup> , Sebastian Sippel<sup>2</sup> , Enikő Székely<sup>3</sup> ,  
Dino Sejdinovic<sup>4</sup> , Emiliano Diaz<sup>1</sup> , Adrián Pérez-Suay<sup>1</sup> , Zhu Li<sup>4</sup>, Miguel Mahecha<sup>5,6</sup>   
and Markus Reichstein<sup>6</sup>

<sup>1</sup> Image Processing Laboratory (IPL), Universitat de València, València, Spain

<sup>2</sup> Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland

<sup>3</sup> Swiss Data Science Center, ETH Zurich and EPFL, Lausanne, Switzerland

<sup>4</sup> Department of Statistics, University of Oxford, Oxford, United Kingdom

<sup>5</sup> Remote Sensing Center for Earth System Research, Leipzig University, Leipzig, Germany

<sup>6</sup> Department of Biogeochemical Integration, Max Planck Institute for Biogeochemistry, Jena, Germany

\* Author to whom any correspondence should be addressed.

E-mail: [jordi.cortes@uv.es](mailto:jordi.cortes@uv.es)

**Keywords:** Earth sciences, detection and attribution, kernel methods, physics-aware machine learning

Supplementary material for this article is available [online](#)

**Abstract**

Process understanding and modeling is at the core of scientific reasoning. Principled *parametric* and mechanistic modeling dominated science and engineering until the recent emergence of machine learning (ML). Despite great success in many areas, ML algorithms in the Earth and climate sciences, and more broadly in physical sciences, are not explicitly designed to be physically-consistent and may, therefore, violate the most basic laws of physics. In this work, motivated by the field of algorithmic fairness, we reconcile data-driven ML with physics modeling by illustrating a *nonparametric* and *nonlinear* physics-aware regression method. By incorporating a dependence-based regularizer, the method leads to models that are consistent with domain knowledge, as reflected by either simulations from physical models or ancillary data. The idea can conversely encourage independence of model predictions with other variables that are known to be uncertain either in their representation or magnitude. The method is computationally efficient and comes with a closed-form analytic solution. Through a consistency-vs-accuracy path diagram, one can assess the consistency between data-driven models and physical models. We demonstrate in three examples on simulations and measurement data in Earth and climate studies that the proposed ML framework allows us to trade-off physical consistency and accuracy.

**1. Introduction**

Physicists and environmental scientists attempt to model systems in a principled way through analytic descriptions that encode scientific understanding and domain expertise of the underlying processes. Conservation laws, physical principles or phenomenological behaviors are generally formalized using mechanistic models and differential equations. Such classical approaches in physics have been, and remain to be, the dominant framework for modeling complex natural Earth and climate phenomena. With the availability of large datasets collected with different remote sensing instruments and, generally, cheap

and widely distributed *in-situ* data collections, the physical modeling paradigm is being complemented, sometimes challenged (and in many cases replaced) by the statistical, machine learning (ML) paradigm, which offers a prior-agnostic approach that does not make direct use of existing scientific knowledge [1–3].

Machine learning models can fit observations very well, but predictions may be physically inconsistent or even implausible. For example, ML models can commit large extrapolation errors, and their predictions can violate fundamental laws like mass or energy conservation [4, 5]. This has been perhaps the most important criticism to ML algorithms, and a relevant reason why, historically, physical modelling

and ML have often been treated as two different fields under very different scientific paradigms (theory-driven versus data-driven). Likewise, there is an ongoing debate around the limitations of traditional methodological frameworks: both about their scientific insight and discovery limits in general [1, 2] and in the geosciences and hydrology in particular [6, 7]. Recently, however, integration of domain knowledge and achievement of physical consistency by teaching ML models about the governing physical laws of the Earth system has been proposed as a principled way to provide strong theoretical constraints on top of the observational ones [4, 8, 9]. The synergy between the two approaches has been gaining attention, with recent approaches including redesigning model's architecture, augmenting the training dataset with simulations, and including physical constraints in the cost function to be optimized [4, 6, 7, 9–15].

The integration of physics in ML models may lead to improved performance and generalization but, more importantly, to improved consistency and credibility of such models. This *hybrid* approach has an interesting regularization interpretation: the inclusion of domain knowledge in the ML model reduces the parameter space to search upon by discarding implausible models. Therefore, physics-aware ML models combat overfitting better, typically become simpler (sparser), and require less training data to achieve similar performance [7, 8, 15, 16]. Physics-aware ML thus leads to enhanced computational efficiency, and constitutes a stepping stone towards the goal of achieving more interpretable ML models [8, 17–20].

Among the many ML models available, kernel methods [21] have shown excellent theoretical properties and practical performance in Earth observation [22] data problems. Kernel methods have been primarily used for classification and regression problems, but also for data clustering, anomaly detection and dimensionality reduction [23]. Kernel methods generalize linear methods easily while still relying on linear algebra operations. The idea is to implicitly map the data into a reproducing kernel Hilbert space (RKHS) [24] where nonlinearities are taken into account, and solve the problem in this new space rather than in the original data space. The solution then typically becomes analytic and only involves simple linear algebra operations on a kernel (similarity) matrix that contains all pairwise similarities between the training data samples.

Our main goal here is to reconcile data-driven models with physics modeling by incorporating physical knowledge in the ML models. More specifically, we propose a nonparametric physics-aware regression method that is based on kernel theory and enables us to understand the role of the physics component from information-theoretic and regularization perspectives. Including prior

knowledge in ML has traditionally been associated with the concept of regularization [20, 25, 26]. Regularizers are typically designed to enforce some desirable features on the model predictions, like smoothness using the  $\ell_2$ -norm on model weights, or directly on the model parameters, like sparsity using an  $\ell_1$ -norm. Recently, other regularizers have been proposed to minimize the sum of all violations of a known physical law [6]. We adopt an alternative regularization approach based on statistical dependence. This approach was introduced for enforcing fairness in model predictions [27], which states that the model predictions should be as statistically independent of predefined sensitive variables as possible. Our hypothesis is that encoding algorithmic fairness and consistency with domain knowledge through regularization play similar roles [28].

Following this idea, our approach achieves (physical) consistency by encouraging the model's predictions to be *dependent* on variables that encode physical knowledge or *independent* from biased information, similarly to the way that fair algorithms ensure that the model's predictions are independent of sensitive or protected variables: income predictions should be arguably independent of gender and race, and in a conceptually similar way, estimates of the forced response in (simulated) global mean temperature should be independent of variations in internal variability in that respective climate change scenario. Therefore, our overarching goal can be defined as:

Learn $\hat{y} = f(x)$ such that $\ y - \hat{y}\ _2^2$ is minimized, and reinforce $\hat{y} \not\perp s$ or $\hat{y} \perp s$ ,
--

that is, learn a function  $f$  that fits well a target variable  $y$  from an input variable  $x$ , and at the same time, either make the predictions  $\hat{y}$  dependent with the ancillary variables  $s$  or statistically independent of them, depending on the problem and application. In this paper, for  $f$  we use a nonparametric regression function and for measuring independence the norm of the cross-covariance operator, both based on kernel methods [23].

Inspired by the fair kernel learning method [27] used in the context of algorithmic fairness, we propose physics-aware kernel learning (PKL). PKL includes a dependence-based regularizer to a given objective function that, depending on the application, may enforce model predictions to resemble a physical model output, a set of simulated data, or additional/supplementary observations (see details in Proposed physics-aware nonparametric regression section). PKL can be trained not only to increase dependence with underlying knowledge but, alternatively, can also be used to ensure predictions are independent of biased/irrelevant information that can arise e.g. from observational errors or anomalous data variability not related to the physical process of interest.

We propose to use the Hilbert–Schmidt independence criterion (HSIC) [29] to measure the dependence between the predictions and physical variables. HSIC excels in capturing all higher order moments of dependence between random variables, is easy to compute and manipulate, and has appealing theoretical properties of convergence to the true dependence measure [29]. The PKL method leads to a closed-form analytic solution. We will show that the PKL method tackles the problem of *physical consistency* and *model-data agreement* in a straightforward manner. The performance of PKL will be illustrated in several examples in Earth and climate sciences. For example, PKL allows us to assess the degree of realism of models in biophysical parameter retrievals, or to detect the effect of external forcing in the climate system while aiming for independence to the exact representation and magnitude of key modes of internal variability. We anticipate that the PKL approach will be of great utility and flexibility for nonparametric data analysis in applied research in general, and in Earth and climate sciences in particular.

## 2. Data collection and pre-processing

### 2.1. Oceanic chlorophyll data

We used the SeaBAM dataset [30, 31], which gathers 919 *in-situ* measurements of chlorophyll concentration around the United States and Europe. The dataset contains *in situ* pigments and remote sensing reflectance measurements ( $R_{rs}$ , [ $\text{sr}^{-1}$ ]) at a set of given wavelengths (412, 443, 490, 510 and 555 nm) that are present in the SeaWiFS ocean color satellite sensor. The chlorophyll concentration values are between 0.019 and 32.79  $\text{mg m}^{-3}$ . Although SeaBAM data originate from various researchers, the variability in the radiometric data is limited. At high Chla concentrations  $CC$  ( $\text{mg m}^{-3}$ ), the dispersion of radiance ratios increases, mostly because of the presence of more optically complex (Case II) waters, that is low values of the ratio of pigment concentration to scattering coefficient. At lowest concentrations the highest  $R(490)/R(555)$  ratios are slightly lower than the theoretical limit for clear natural waters. More information about the data can be obtained from [SEABAM](#), and an extensive analysis in [31]. In addition to the observational data, we used several models to guide PKL. Morel1 and CalCOFI 2-band linear are described by  $s = 10^{a_0 + a_1 \kappa}$ , while OC2/OC4 models follow  $s = a_0 + 10^{a_1 + a_2 \kappa + a_3 \kappa^2 + a_4 \kappa^3}$ . The ratio  $\kappa$  depends on the physical model used [31]: the Morel1 model uses  $\kappa = \log(R(443)/R(555))$ , the CalCOFI and OC2 models use  $\kappa = \log(R(490)/R(555))$ , while the OC4 model uses  $\kappa = \log(\max\{R(443), R(490), R(510)\}/R(555))$ . The goal here is to predict the concentrations from reflectance measurements while being consistent with these parametric models that encapsulate some domain knowledge.

### 2.2. Hyperspectral and vegetation *in situ* data

We collected hyperspectral data from the CHRIS sensor as well as *in situ* measurements of chlorophyll content (Chla), leaf area index (LAI) and fractional vegetation cover (fCOVER) [32]. LAI is a dimensionless quantity that characterizes plant canopies. It is defined as the one-sided green leaf area per unit ground surface area in broadleaf canopies. fCOVER corresponds to the fraction of ground covered by green vegetation, and in practice it quantifies the spatial extent of the vegetation. It is indeed independent from the illumination direction and sensitive to the vegetation amount and derived from the LAI plus a number of structural parameters of the canopy. The data were obtained during the SPARC-2003 (SPeetra bARrax Campaign) and SPARC-2004 campaigns in Barrax, Spain. The region consists of approximately 65% dry land and 35% irrigated land. Green LAI was derived from canopy measurements made with a LiCor LAI-2000 digital analyzer. Each elementary sampling unit (ESU) was assigned to a LAI value, which was obtained by the average of 24 measures (8 data readings  $\times$  3 replications) [33]. fCOVER was estimated from ground measurements using hemispherical photographs taken with a digital camera with a fish-eye lens. The final fCOVER estimate for each ESU was calculated as the average of twelve measurements. In total, nine crop types (garlic, alfalfa, onion, sunflower, corn, potato, sugar beet, vineyard and wheat) were sampled, with field-measured values of LAI that vary between 0.4 and 6.3, Chla between 2 and 55  $\mu\text{g cm}^{-2}$  and fCOVER between 0 and 1. Additionally, 30 random bare soil spectra with a biophysical (Chla, LAI, fCOVER) value of zero were added to broaden the dataset to non-vegetated samples. Concurrently, we used CHRIS images Mode 1 (62 spectral bands, 34 m spatial resolution at nadir). The images were geometrically and atmospherically corrected. A total of  $n = 136$  data points in a 62-dimensional space were thus used to fit a PKL model. The goal here was to estimate LAI while being consistent with fCOVER, and estimate chlorophyll content while being consistent to LAI.

### 2.3. Internal variability and forced variability data

The US CLIVAR Working Group on large ensembles (LEs) contains a data archive of initial-condition LEs conducted with different climate models run within their CMIP5 setup. From the seven climate models available, we selected three different ones: CanESM2 for training, CSIRO-Mk3-6-0 for validation, and CESM1-CAM5 for testing. We emphasize that different train/validation/test splits are possible (and should be tested in real-world applications), but here we only show an illustration which is why we present this setup and given that results are robust for alternative choices. Model simulations for all three contained the period between 1950 and 2100

with historical and RCP 8.5 forcing conditions and incorporated more than 30 different ensemble members [34]. All model runs were defined with a spatial regridded world map of  $5^\circ \times 5^\circ$  resolution (in total  $d = 2592$  grid points). The aggregated ensemble of each model, and all the grid points, defined the predictors matrix. We split it into a training, validation and test sets. We selected the spatially explicit simulated monthly near-surface air temperature (TAS) as our variable of interest (i.e. predictors). The ENSO internal variability is captured by the Niño3.4 index, and is taken from the climate variability diagnostics package for LEs developed by NCAR's Climate Analysis section [35]. Our goal is to predict the forced climate response from a single ensemble member, and the 'true' forced response was extracted as the average across the full ensemble, which is a standard procedure in the field. The predictors and metrics used were taken as the DJF seasonal average and standardized to zero mean and unit standard deviation.

### 3. Proposed physics-aware nonparameteric regression

We are given a set of inputs,  $\mathbf{x}_i \in \mathcal{X}$ , and the corresponding targets,  $y_i \in \mathcal{Y}$ , for  $i = 1, \dots, n$ . Furthermore, we define  $\mathbf{s}_i \in \mathcal{S}$  the set of variables to which we want to emphasize the (in)dependence. These will be referred to as the *ancillary* variables. We take  $\mathbf{x}_i$  to be an i.i.d. sample from an  $\mathcal{X}$ -valued random variable  $\mathbf{x}$ , and similarly for  $\mathbf{s}$ . For simplicity, we will assume that the inputs are vectorial, i.e.  $\mathbf{x}_i \in \mathbb{R}^{d \times 1}$ ,  $\mathbf{s}_i \in \mathbb{R}^{q \times 1}$  and that the targets are scalars, i.e.  $y_i \in \mathbb{R}$ , but the exposition can be trivially extended to non-Euclidean or structured domains which admit positive definite kernel functions. We let  $\mathbf{X} \in \mathbb{R}^{n \times d}$  denote the matrix of  $n$  observed inputs corresponding to  $d$  explanatory covariates,  $\mathbf{S} \in \mathbb{R}^{n \times q}$  denotes the matrix of  $n$  observations of  $q$  ancillary variables,  $\mathbf{y} \in \mathbb{R}^{n \times 1}$  denotes the vector of observed targets, which we could assume are distorted with biases or noise, and  $\hat{\mathbf{y}}$  is the prediction.

Fitting a consistent-regularized model  $f_* \in \mathcal{H}$  for some hypothesis class  $\mathcal{H}$  reduces to optimizing a regularized empirical risk functional [27, 36]:

$$f_* = \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n V(f(\mathbf{x}_i), y_i) + \lambda \Omega(f) + \mu I(f(\mathbf{x}), \mathbf{s}) \right\},$$

where  $V$  is the loss function,  $\Omega$  acts as an overfitting/complexity penalty on  $f$ , and  $I$  measures the statistical (in)dependence between the model  $f$  and the ancillary variables, the latter depending on the sign that accompanies  $I$ . The two regularization parameters  $\lambda \in [0, \infty)$  and  $\mu \in (-\infty, +\infty)$  control smoothness and consistency of the solution, respectively. Note that the sign of  $\mu$  forces dependence or independence. By setting  $\mu = 0$ , the solution of the standard

kernel ridge regression model is obtained. This does not mean that the solution is inconsistent but one can find a more consistent model, and thus conflicting less with ancillary information, by changing  $\mu$ . As  $\mu \neq 0$  implies a trade-off between different objectives, selecting an optimal value is subject on the application [37]. On another note, one should be aware that the problem could be ill-conditioned for high negative values of  $\mu$ .

For the consistency penalization term,  $I$ , we adopted the HSIC [29] between the predicted response  $\mathbf{f} = [f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)]$  and the ancillary variable  $\mathbf{s} = [s_1, \dots, s_n]$ . The HSIC measures independence between random variables  $\mathbf{f}$  and  $\mathbf{s}$ . Given the dataset  $\mathcal{D}$  with  $n$  samples drawn from the joint distribution  $P(\mathbf{f}, \mathbf{s})$ , an empirical estimator of HSIC is defined as [29]:

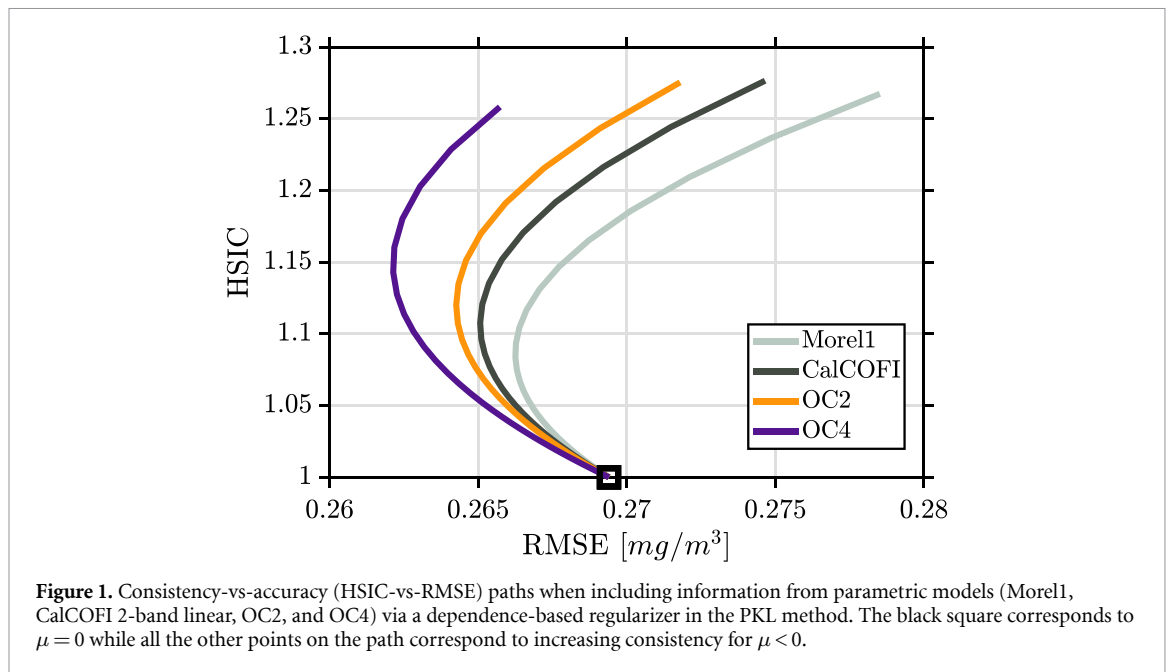
$$\widehat{\text{HSIC}}_{k,l}(\mathbf{f}, \mathbf{s}) = \frac{1}{n^2} \text{Tr}(\mathbf{KHLH}),$$

where  $\mathbf{K}$  and  $\mathbf{L}$  are the kernel matrices computed on observations  $\{f(\mathbf{x}_i)\}_{i=1}^n$  and  $\{s_i\}_{i=1}^n$  using kernel functions  $k$  and  $l$ , respectively, and  $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^\top$  has the role of centering the data in the feature space. For a broad family of kernels  $k$  and  $l$ , the population HSIC equals 0 if and only if the two involved variables are statistically independent [29]. With appropriate choices of kernels  $k$  and  $l$  for the input data  $\mathbf{X}$  and the ancillary variable(s)  $\mathbf{S}$ , respectively, the HSIC regularizer captures all types of statistical dependence between  $f$  and the ancillary variable  $\mathbf{s}$ .

In this work we use the HSIC regularization in combination with kernel ridge regression as the model function class, which leads to a closed-form (analytic) solution. This regularization can be incorporated in other ML models, like neural networks and Gaussian processes. More information on the kernel physics-aware kernel regression (PKL) solution used in this work, and connections to other ML models can be found in the appendices. An illustration of the PKL method for a toy example can also be found in the supplementary material S1 (available online at [stacks.iop.org/ERL/17/054034/mmedia](https://stacks.iop.org/ERL/17/054034/mmedia)).

### 4. Results and discussion

We describe three cases to illustrate the capabilities of PKL to encode physical knowledge about the system under consideration and to assess consistency between physical knowledge and the data. The PKL solutions are summarized in the form of an easy to understand consistency-vs-accuracy path diagram, i.e. HSIC-vs-RMSE path, that describes the relationship between the degree of consistency with physical knowledge and the accuracy of the regression model. This relationship is described as a function of the amount of dependence-based regularization that is captured by a hyperparameter  $\mu$ ,



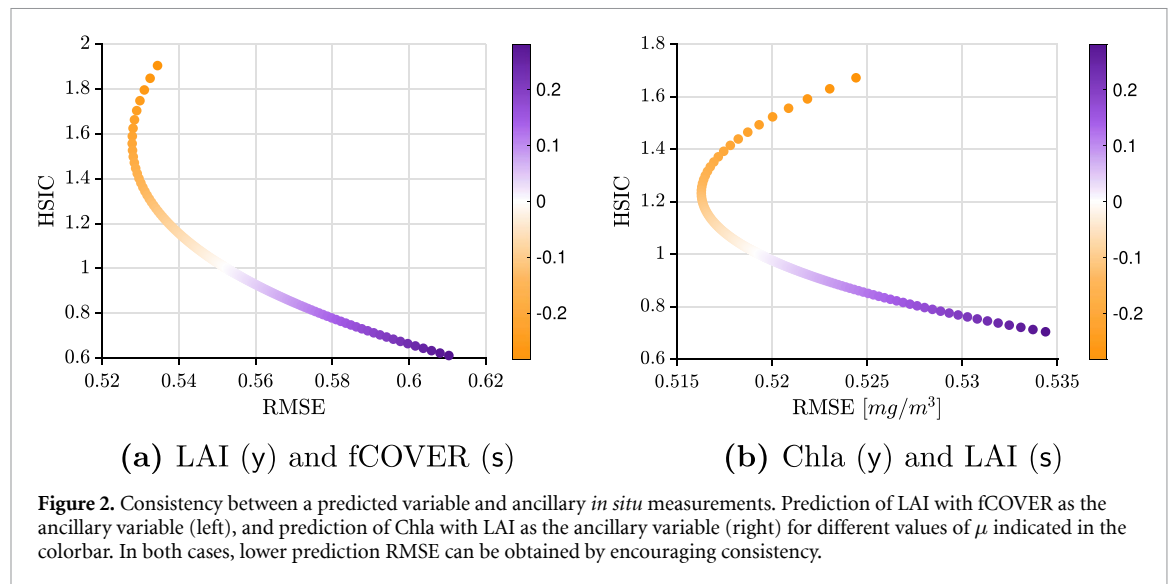
which varies between  $-\infty$  (dependence) and  $+\infty$  (independence). By observing how the path diagram changes upon encouraging (or alternatively reducing) dependence, one obtains an indication of how physical variables and model predictions are connected. This analysis allows us to assess the agreement between model predictions and a set of physical variables by comparing their corresponding path diagrams. Considering that HSIC takes values between zero and infinity, the corresponding HSIC ( $\mu$ ) values are then scaled by the HSIC ( $\mu = 0$ ) value to allow comparison across multiple consistency-vs-accuracy paths: the normalized HSIC is equal to 1 when  $\mu = 0$ , and corresponds to the standard kernel regression solution. In what follows we will use  $x$  for the input,  $y$  for the target and  $s$  for the ancillary variables (see proposed physics-aware nonparametric regression section).

#### 4.1. Consistency with models for biophysical parameter estimation

A classical problem in remote sensing and geosciences involves estimation of biophysical parameters of interest from remote sensing (satellite) observations. We are given multidimensional measurements  $\mathbf{x} \in \mathbb{R}^d$ , acquired from a satellite sensor at  $d$  different wavelengths, from which we want to estimate a parameter of interest  $y \in \mathbb{R}$ . This can be done using satellite and *in-situ* measurement pairs  $(\mathbf{x}, y)$ . Traditionally, the community has designed a great many empirical, *parametric* models for estimation. We aim to use the PKL method to solve the fitting problem and, more importantly, to assess the most appropriate parametric model to rely on by studying the consistency-vs-accuracy path diagrams.

We first illustrate the performance of the PKL model for the estimation of *in-situ* chlorophyll

concentration from multispectral reflectance images. The two measurements are subject to high levels of uncertainty, owing both to the difficulties in ground-truth data acquisition, and the noise inherent in satellite-obtained data. Moreover, there is commonly a time mismatch between the acquired image and the recorded *in-situ* measurements, which is critical, for instance, for coastal water monitoring. We use the SeaBAM dataset [30, 31] (see details in Data section), and apply PKL to model ocean chlorophyll concentration  $y$  from radiances  $\mathbf{x} := [R(\lambda_1), \dots, R(\lambda_d)] \in \mathbb{R}^d$  at a set of given wavelengths,  $\lambda_i$ , while encouraging consistency with four standard parametric models (Morel1, CalCOFI 2-band linear, OC2 and OC4) [31] in a set of four different experiments, one for each model as the ancillary variable  $s$ . Results are shown in figure 1. PKL solves the fitting problem with improved accuracy, i.e. a lower error (RMSE) can be achieved compared to standard kernel ridge regression ( $\mu = 0$ ), and allows to quantify the quality of the different parametric models considered through their respective consistency-vs-accuracy paths. Encouraging consistency with more recent models, such as OC2 and OC4, leads to a lower PKL prediction error (lower RMSE) compared to older models, such as Morel1 and CalCOFI, meaning that they fit the data better (i.e. the consistency between the data and the model is higher). When increasing dependence with each parametric model, the HSIC regularizer begins to dominate the training cost function and similarity of the target variable with the ancillary variable is reinforced. This happens until a certain turning point in  $\mu$  where too much dependency leads to an increased error. In summary, this simple example shows how PKL may benefit the estimation of Earth system parameters such as the *in-situ* chlorophyll concentration from remote sensing by enforcing



consistency with well-known parametric models that encapsulate domain knowledge.

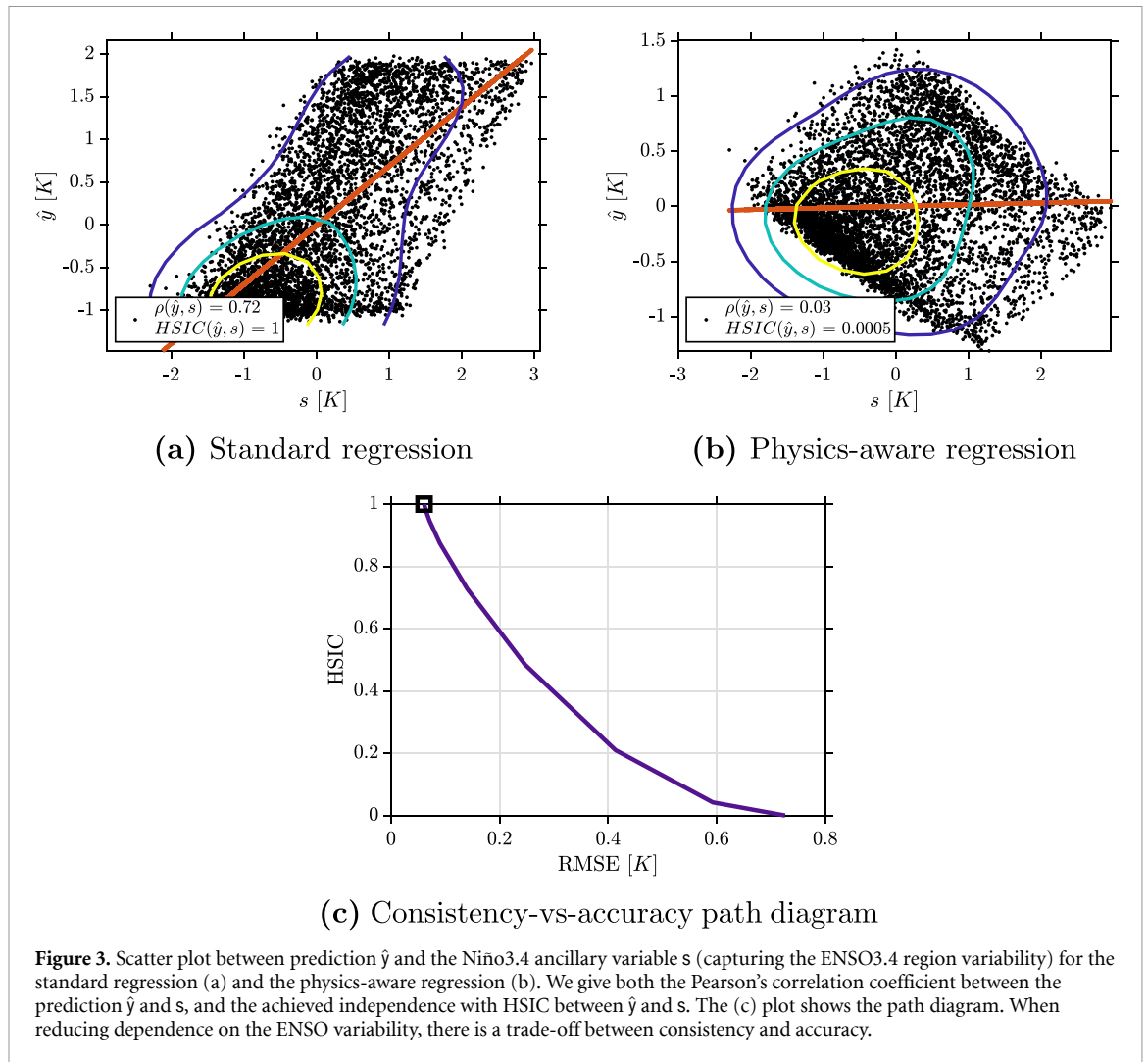
#### 4.2. Consistency with ancillary *in situ* data

Very often one does not have access to a parametric model as in the previous example, but to a set of ancillary observational data that the predictions should be consistent with. This is a standard case in remote sensing where some variables are coupled, e.g. greenness and chlorophyll content of the vegetation. We illustrate the use of the PKL in two cases: (1) to estimate leaf area index ( $y := \text{LAI}$ ) while being consistent with the fraction of green vegetation cover ( $s := \text{fCOVER}$ ), and (2) to predict chlorophyll-a concentration ( $y := \text{Chla}$ ), while being consistent with the leaf area index ( $s := \text{LAI}$ ). We use data from a terrestrial campaign where hyperspectral images and the aforementioned biophysical parameters were measured (see Data section for details). Figure 2 shows the joint evolution of HSIC and RMSE. The color intensity along the paths represents the increasing (decreasing) value of the consistency hyperparameter  $\mu$  as it goes from reducing consistency with the ancillary variable (positive) to increasing consistency (negative). Optimal solutions, i.e. predictions with lower RMSE and higher HSIC to the ancillary variable, are obtained for negative values of  $\mu$ , meaning that increasing the consistency between the variables (LAI-fCOVER and Chla-LAI) also helps in reducing the RMSE of the prediction. Similarly to the example of parametric models, higher consistency allows for higher accuracy (lower error) up to certain values starting from which a trade-off appears between consistency and accuracy. This example illustrated that, while many accurate solutions are possible when retrieving biophysical parameters with ML, one can achieve an optimal solution that is both accurate and ensures consistent results across variables.

#### 4.3. Learning patterns of forced warming under uncertain climate variability

Separating forced and internal variability is a key challenge in climate science [34, 38, 39], and in particular in the context of detection and attribution (D&A) of externally forced climate change [40, 41]. D&A typically uses fingerprints that encapsulate the structure of forced warming from model simulations into a spatial or spatiotemporal pattern [40–42]. Observations and climate model control simulations are then projected onto those fingerprints to assess whether a signal, such as multidecadal temperature trends at the global or continental scale, exceed internal climate variations [42].

In recent research, forced climate signals have been extracted from observations via signal-to-noise optimized fingerprints from climate models obtained through statistical learning [39, 43–45], thus building on ideas from traditional D&A. Statistical learning algorithms in this context aim to minimize the influence of internal variability and model disagreement to extract a forced signal [46]. However, patterns of internal variability (such as El Niño Southern Oscillation, ENSO) may nonetheless project onto the fingerprint, for example, if the fingerprint is derived from a given model (or a given set of models) but then applied to an unseen test model (or observations) with a potentially different representation of key modes of internal variability, such as ENSO. This phenomenon may lead to an inaccurate extraction of forced signals from models or observations, and could lead to over- or under-confidence in D&A statements if models systematically under- or overestimate the magnitude of internal variability [44, 47]. Hence, robustness in fingerprint extraction with respect to model structural differences, or potential systematic biases in the models' representation of internal variability or of forced patterns, is an important issue



**Figure 3.** Scatter plot between prediction  $\hat{y}$  and the Niño3.4 ancillary variable  $s$  (capturing the ENSO3.4 region variability) for the standard regression (a) and the physics-aware regression (b). We give both the Pearson's correlation coefficient between the prediction  $\hat{y}$  and  $s$ , and the achieved independence with HSIC between  $\hat{y}$  and  $s$ . The (c) plot shows the path diagram. When reducing dependence on the ENSO variability, there is a trade-off between consistency and accuracy.

that still remains a key uncertainty. This issue is related to distributional robustness in statistics [48] and transfer learning in ML [49]. For example, climate models simulate sea surface temperature patterns and warming that are broadly consistent with observations on a global scale [50], and also capture key modes of internal variability such as ENSO [51]. However, models differ (1) in the time scales of irregular ENSO fluctuations, (2) in the characteristic patterns of ENSO-related climate variability, and (3) in how ENSO variability is projected to evolve in a warming climate [51]. Moreover, differences across models in the forced warming and variability in multidecadal trends are particularly pronounced in the equatorial Pacific [50, 52]. While historical model simulations show generally warming sea surface temperatures in the equatorial Pacific (e.g. ENSO3.4 region:  $[-170^\circ \text{E}, -120^\circ \text{E}]$  and  $[-5^\circ \text{N}, 5^\circ \text{N}]$ ), observations show only very modest warming in this region over multi-decadal time scales [52]. It is currently unclear whether this discrepancy is due to an unusual realization of the observed climate (i.e. internal variability), or whether the models show systematic biases in the forced response compared to observations [50].

Here, we illustrate the idea of estimating the forced climate response from individual ensemble members (similar to e.g. [44, 46]) but in a way that takes into account the ENSO-related variability by reducing consistency to the ENSO3.4 region variability expressed by the Niño3.4 index. Reducing consistency in this context thus implies that ENSO-related variability does not project onto the fingerprint (i.e. the prediction of the forced response should be independent of the ENSO3.4 region variability), and hence our goal is to attain improved robustness in the context of uncertainties related to variability in the equatorial Pacific.

For this purpose, we employ the multi-model large ensemble archive [34] and the climate variability diagnostics package [35] (more details in Data section). Winter (December–January–February, DJF) seasonal average values of the forced response in each climate model are calculated as the ensemble average of multiple model runs that differ from each other by a small perturbation. By taking the average of the ensemble, the small oscillations between models that are attributed to internal variability are smoothed out and the end product reflects the forced climate

response in the absence of variability [35]. These values constitute the target variable  $y$  of our prediction. The statistical model is based on DJF anomalies in individual ensemble members as predictor variables  $x$ . We run the PKL with Niño3.4 as the ancillary variable  $s$  that we want the prediction to be independent to (i.e. PKL is run with positive increasing values of the consistency hyperparameter,  $\mu > 0$ ). For illustration, we employ a different climate model for training (CanESM2), validation (CSIRO-Mk3-6-0) and testing (CESM1-CAM5).

Results are presented in figure 3. Without reducing consistency ( $\mu = 0$ ), the prediction  $\hat{y}$  and Niño3.4 are fairly correlated (Pearson's correlation  $\rho = -0.72$ ) and dependent from the prediction for the unseen test model (figure 3(a)). This implies that ENSO-related uncertainties or potential systematic biases may alias into the prediction of the forced response for an unseen model or in observations. However, if Niño3.4 is taken as the ancillary variable and upon reducing consistency,  $\mu > 0$ , the prediction becomes more decorrelated ( $\rho(\hat{y}, s) = 0.03$ ) and independent (HSIC  $(\hat{y}, s) = 0.0005$ ) of the variability introduced by ENSO3.4 (figure 3(b)). The results imply that our forced response estimate would be more robust to ENSO-related uncertainties with an appropriate magnitude of  $\mu$  across different climate models or between models and observations. An illustration of the effect the independence constraint has for the prediction can be seen in the supplementary material S2.

In summary, we have illustrated that estimating a climate signal (such as the forced response) from climate variables can be learnt in a way that is 'consistent' to certain known model structural uncertainties such as the representation of internal variability and warming in the equatorial Pacific that determine ENSO-related variability in the climate system.

## 5. Conclusions

Machine learning models have been traditionally considered black box models where interpretability of the decisions is hidden behind a complex architecture. ML models do not necessarily carry physical meaning, even if generally accurate, and hence they can produce non-physical or even nonsensical predictions. A robust trustworthy model should be in accordance with the known and agreed rules of the physical world. This dichotomy between how humans encode knowledge mathematically in mechanistic models and how ML models encode knowledge through their expression learnt from data has been the reason for longstanding debates in the last century: data-driven versus model-driven science.

In the last decade, performing accurate predictions has replaced process understanding in many applied areas of science and engineering. In these areas it is believed that *data* is the only needed

regularizer of the model class. The problem, however, comes in cases of data-limited regimes, model misspecification, or omitted-variable bias. In all these cases, the ML model is an incomplete and often simplistic representation of reality, and generally attributes the effect of the missing variables to the estimated effects of the included variables. Besides, one can achieve similarly accurate results with completely different models; that is the issue of equifinality or non-identifiability. In recent years, many efforts have been conducted to develop ML algorithms that are more transparent, accountable, interpretable and consistent with first principles. However, developing hybrid ML models that include expert/domain knowledge (in the form of physics rules, parameterization and constraints) have been only marginally treated in the recent literature, for example, using deep neural networks, kernel methods and Gaussian processes [4, 6, 9–15]. The field is interesting in broad terms and generally applicable to all domains where data and models coexist.

We proposed a general methodology based on kernels to combine two types of knowledge priors: smoothness of the function class and consistency of the predicted target with ancillary variables, with particular focus on Earth and climate sciences. Motivated from the algorithmic fairness literature, we introduced a physics-aware kernel-based model that exploits ancillary information or outputs from mechanistic models to *regularize* pure data-driven ML based on scientific knowledge. The PKL regression model includes a regularizer that measures the dependence of the learned function with physical variables, ancillary data or model simulations. The selection of a kernel-based regularizer based on HSIC leads to a simple and analytic solution where a probabilistic model interpretation is also amenable. HSIC ensures fast (exponential) convergence rates of the population measure to the population quantity as the sample size grows. PKL generalizes both linear and nonlinear kernel-based regression models, it is easy to implement and inherits all properties of the kernel methods treatment. For instance, recent advances in kernel and Gaussian processes modeling could be included in our predictive function, e.g. sparse formulations for improved computational scalability, deep models for improved expressive functions, or machines to learn ordinary/partial differential equations and control [10, 11, 15, 53]. Likewise, the HSIC regularizer could be used in other ML models like neural networks. See appendices for more details on the derivation of PKL, a GP probabilistic interpretation of the HSIC regularizer and the use in standard feedforward neural networks. The present work focuses on kernel methods for their simplicity, since solutions are analytically available. Nevertheless, adding a physics-aware regularizer in the form of HSIC can be included in the loss function of other types of ML methods. The PKL model is fully



functional, and being implemented as an open-source software tool, it allows to reuse current and upcoming kernel models and tools.

We anticipate that the PKL will constitute a convenient, useful and flexible tool for nonparametric data analysis in applied research where data, process understanding, principled models, and simulations are generally available. This is the general case in Earth sciences, but also for climate science studies where statistical learning is often adversely affected by uncertainties in regional variability characteristics, disagreement between models and structural uncertainty, and systematic biases.

### Data availability statement

The data that support the findings of this study will be openly available following an embargo at the following URL/DOI [54]: <https://github.com/IPL-UV/PKL>. Data will be available from 18 April 2022.

### Acknowledgments

G C V would like to acknowledge the support from the European Research Council (ERC) under the ERC Consolidator Grant 2014 project SEDAL (Grant Agreement No. 647423). G C V and M R thank the support of ERC Synergy Grant USMILE (Grant Agreement No. 855187). The work of S S and E S was partly funded by the Swiss Data Science Center within the project 'Data Science-informed attribution of changes in the Hydrological cycle' (DASH, C17-01). G C V, S S, M M and M R thank the support by the European Union's Horizon 2020 research and innovation programme within project 'XAIDA: Extreme Events—Artificial Intelligence for Detection and Attribution', under Grant Agreement No. 101003469. This paper was partly inspired by discussions on the first ELLIS workshop on 'Machine Learning for Earth and Climate Sciences' 2–4 March 2020, just a week before the European lockdown.

## Appendix

### Consistency regularization framework

The PKL functional for the function class  $f$  used in this work, is defined as:

$$f_* = \arg \min_{f \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^n V(f(\mathbf{x}_i), y_i) + \frac{\lambda}{n} \|f\|_{\mathcal{H}_k}^2 + \mu \widehat{\text{HSIC}}_{k,l}(f(\mathbf{x}), \mathbf{s}) \right\},$$

where we adopted the reproducing RKHS  $\mathcal{H}_k$  as the hypothesis class. Given that the selected consistency

penalty term, HSIC, only depends on the unknown function  $f$  through its evaluations at the training inputs  $\{\mathbf{x}_i\}$ , direct application of the Representer's theorem [55] tells us that the optimal solution can be written as a linear combination of kernel function evaluations  $f = \sum_{i=1}^n \alpha_i k(\cdot, \mathbf{x}_i)$ . Hence, we obtain the so-called *dual* problem:

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \|\mathbf{K}\alpha - \mathbf{y}\|_2^2 + \frac{\lambda}{n} \alpha^\top \mathbf{K} \alpha + \frac{\mu}{n^2} \alpha^\top \mathbf{H} \mathbf{L} \mathbf{H} \mathbf{K} \alpha \right\},$$

which can now be solved for  $\alpha$  directly. In the case of the squared  $L_2$  norm for  $V := \|f(\mathbf{x}) - \mathbf{y}\|_2^2$ , the above dual problem has an analytic solution [27]:

$$\alpha = \left( \mathbf{K} + \frac{\lambda}{n} \mathbf{I} + \frac{\mu}{n^2} \mathbf{H} \mathbf{L} \mathbf{H} \mathbf{K} \right)^{-1} \mathbf{y}.$$

The dependence estimator is sensitive to the hyperparameters of the kernel functions  $k$  and  $l$ . To avoid variations in HSIC exclusively due to changes in these parameter values, some form of normalization is needed. Nevertheless, one cannot normalize both kernels arbitrarily since the optimization problem would become nonlinear with  $\alpha$  and would not admit a closed-form solution anymore. As the parameters from  $\mathbf{K}$  are adjusted to the data during the optimization process, one could partially normalize the cross-covariance on  $\mathbf{L}$ . We introduce a modification inspired by the normalized version of HSIC called NOCCO [56]. This modification simply replaces  $\mathbf{H} \mathbf{L} \mathbf{H}$  with  $\mathbf{H} \mathbf{L} \mathbf{H} (\mathbf{H} \mathbf{L} \mathbf{H} + n\epsilon \mathbf{I})^{-1}$ , where  $\epsilon$  is a regularization parameter used in the same way as [56]. The solution then becomes:

$$\alpha = \left( \mathbf{K} + \frac{\lambda}{n} \mathbf{I} + \frac{\mu}{n^2} \mathbf{H} \mathbf{L} \mathbf{H} (\mathbf{H} \mathbf{L} \mathbf{H} + n\epsilon \mathbf{I})^{-1} \mathbf{K} \right)^{-1} \mathbf{y}.$$

In all our experiments, we used the squared exponential (SE) kernel for both  $k$  and  $l$ , that is e.g.  $k(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2 / (2\sigma^2))$ , which captures sample similarity well in most of the problems, and only one hyperparameter  $\sigma$  needs to be tuned. For the standard kernel ridge regression (KRR) we implemented a standard cross-validation scheme to determine the ridge regression parameters,  $\sigma$  and  $\lambda$ , setting the squared error loss as the metric to minimize. For the ancillary variable we set the SE hyperparameter to the average Euclidean distance between all points, which is a commonly used heuristic in the kernel methods literature [23, 57]. A much larger sigma would make HSIC approximate a linear dependence estimate, while a much smaller lengthscale  $\sigma$  would make HSIC insensitive to dependence [29].

## A probabilistic treatment: the Gaussian process approach

Using HSIC in a kernel-based regression framework can be actually seen as a modified Gaussian process prior. From a Gaussian Process (GP) perspective, the prior covariance corresponds to a posterior covariance for meta-observations encouraging that the predictive function  $f$  remains constant as the ancillary variables vary. Let us assume that the loss corresponds to the negative conditional log-likelihood in some probabilistic model, i.e. that  $V(f(\mathbf{x}_i), y_i) = -\log p(y_i|f(\mathbf{x}_i))$ , which is true for a wide class of loss functions. We write  $V(\mathbf{y}, \mathbf{f}) = -\sum_{i=1}^n \log p(y_i|f(\mathbf{x}_i))$  to denote the rescaled conditional negative log-likelihood. Consider now using explicit feature mapping  $\mathbf{x}_i \mapsto \phi(\mathbf{x}_i)$  and denoting the feature matrix by  $\Phi$ , we have  $\mathbf{f} = \Phi\beta$  and recast optimization as:

$$\min_{\beta \in \mathbb{R}^m} \left\{ \frac{1}{\lambda} V(\mathbf{y}, \Phi\beta) + \beta^\top \beta + \delta \beta^\top \Phi^\top \mathbf{H} \mathbf{L} \mathbf{H} \Phi \beta \right\},$$

where  $\beta$  play the same role as  $\alpha$  in the PKL, and we defined  $\delta = \mu/\lambda n$  for convenience. The two regularization terms correspond, up to an additive constant, to a negative log-prior of  $\beta \sim \mathcal{N}\left(0, \left(\mathbf{I} + \delta \Phi^\top \mathbf{H} \mathbf{L} \mathbf{H} \Phi\right)^{-1}\right)$ , which in turn gives a prior on the evaluations:

$$\mathbf{f} \sim \mathcal{N}\left(0, \Phi \left(\mathbf{I} + \delta \Phi^\top \mathbf{H} \mathbf{L} \mathbf{H} \Phi\right)^{-1} \Phi^\top\right).$$

By directly applying the Woodbury-Morrison formula, the covariance matrix in this prior becomes  $(\mathbf{K}^{-1} + \delta \mathbf{H} \mathbf{L} \mathbf{H})^{-1}$ , compared to  $\mathbf{K}$  in the standard GP case. Thus, adding an HSIC regularizer corresponds to modifying the prior on function evaluations  $\mathbf{f}$ . The posterior mode in a Bayesian model using a modified GP prior becomes:

$$f \sim \mathcal{GP}\left(0, k(\cdot, \cdot) - k_{\mathbf{X}}^\top (\mathbf{K} \mathbf{H} \mathbf{L} \mathbf{H} + \delta^{-1} \mathbf{I})^{-1} \mathbf{H} \mathbf{L} \mathbf{H} k_{\mathbf{X}}\right).$$

where  $k_{\mathbf{X}} = [k(\cdot, \mathbf{x}_1), \dots, k(\cdot, \mathbf{x}_n)]^\top$ , for any training set  $\{\mathbf{x}_i\}_{i=1}^n$ . Treating the learning problem in the GP framework actually allows to derive uncertainty estimates and hyperparameter tuning using marginal likelihood maximization.

## Neural network training with the physics-dependence loss

Neural networks can also benefit from the new loss including the dependence term. Including the HSIC term in standard backpropagation leads to simple updating rules for training neural networks. Let us denote a feedforward neural network function  $\hat{\mathbf{y}} = \mathcal{F}(\mathbf{X}; \mathbf{W})$ , parameterized by a set of

weights  $\mathbf{W}$ . The backpropagation algorithm uses gradient descent to adjust model weights iteratively  $\mathbf{W}[t] \leftarrow \mathbf{W}[t-1] + \eta \nabla J_{\mathbf{W}}$ , where  $J = \|\mathbf{e}\|$  is the loss (cost, energy) function that depends on training error  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  and  $\eta > 0$  is the learning rate. Including the physics-dependence loss in a neural network training is straightforward, and only involves replacing the output error  $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$  with  $\mathbf{e}' = \mathbf{y} - (\lambda \mathbf{I} + \mu \tilde{\mathbf{S}} \tilde{\mathbf{S}}^\top) \hat{\mathbf{y}}$ , which intuitively trades the training error for physics consistency. The standard backpropagation algorithm can be applied to the new error function,  $J_{\mathbf{W}}' := \|\mathbf{e}'\|^2$ . Alternatively one could use other training algorithms like Adam or automatic differentiation if the interest is to learn the dependence measure parameters end-to-end. The algorithm allows us to train large scale problems while encoding physical consistency concepts easily.

## ORCID iDs


Jordi Cortés-Andrés  <https://orcid.org/0000-0002-9344-8191>

Gustau Camps-Valls  <https://orcid.org/0000-0003-1683-2138>

Sebastian Sippel  <https://orcid.org/0000-0002-4510-4458>

Enikő Székely  <https://orcid.org/0000-0001-5710-9814>

Dino Sejdinovic  <https://orcid.org/0000-0001-5547-9213>

Emiliano Diaz  <https://orcid.org/0000-0001-8410-6635>

Adrián Pérez-Suay  <https://orcid.org/0000-0002-8258-4454>

Miguel Mahecha  <https://orcid.org/0000-0003-3031-613X>

Markus Reichstein  <https://orcid.org/0000-0001-5736-1112>

## References

- [1] Halevy A, Norvig P and Pereira F 2009 The unreasonable effectiveness of data *IEEE Intell. Syst.* **24** 8–12
- [2] Lipton Z C 2018 The mythos of model interpretability *Queue* **16** 31–57
- [3] Emmanuel de B, Pajot A and Gallinari P 2019 Deep learning for physical processes: incorporating prior scientific knowledge *J. Statist. Mech.: Theory and Experiment* **2019** 12
- [4] Reichstein M, Camps-Valls G, Stevens B, Denzler J, Carvalhais N, Jung M and Prabhat 2019 Deep learning and process understanding for data-driven Earth system science *Nature* **566** 195–204
- [5] Marcus G 2018 Deep learning: a critical appraisal (arXiv:1801.00631)
- [6] Karpatne A, Atluri G, Faghmous J H, Steinbach M, Banerjee A, Ganguly A, Shekhar S, Samatova N and Kumar V 2017 Theory-guided data science: a new paradigm for scientific discovery from data *IEEE Trans. Knowl. Data Eng.* **29** 2318–31

- [7] Zhao W Li, Gentine P, Reichstein M, Zhang Y, Zhou S, Wen Y, Lin C, Li Xi and Qiu G Y 2019 Physics-constrained machine learning of evapotranspiration *Geophys. Res. Lett.* **46** 14496–507
- [8] Willard J, Jia X, Shaoming X, Steinbach M and Kumar V 2020 Integrating physics-based modeling with machine learning: a survey (arXiv:2003.04919)
- [9] Pawar S, San O, Aksoylu B, Rasheed A and Kvamsdal T 2021 Physics guided machine learning using simplified theories *Phys. Fluids* **33** 011701
- [10] Svendsen H D, Martino L, Campos-Taberner M, García-Haro J and Camps-Valls G 2017 Joint gaussian processes for biophysical parameter retrieval *IEEE Trans. Geosci. Remote Sens.* **1** 1718–27
- [11] Camps-Valls G, Martino L, Svendsen D H, Campos-Taberner M, Mu noz-Mari J, Laparra V, Luengo D and García-Haro F J 2018 Physics-aware gaussian processes in remote sensing *Appl. Soft Comput.* **68** 69–82
- [12] Kashinath K, Albert A, Wang R, Mustafa M and Rose Y 2019 Physics-informed spatio-temporal deep learning models *Bull. Am. Phys. Soc.* **64** 13
- [13] Jin-Long W, Xiao H and Paterson E 2018 Physics-informed machine learning approach for augmenting turbulence models: a comprehensive framework *Phys. Rev. Fluids* **3** 074602
- [14] Raissi M, Perdikaris P and Karniadakis G E 2019 Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations *J. Comput. Phys.* **378** 686–707
- [15] Svendsen D H, Piles M, Mu noz-Mari J, Luengo D, Martino L and Camps-Valls G 2021 Integrating domain knowledge in data-driven earth observation with process convolutions *IEEE Trans. Geosci. Remote Sens.* **48** 1–14
- [16] Stewart R and Ermon S 2016 Label-free supervision of neural networks with physics and domain knowledge AAAI'17: *Proc. of the 31st AAAI Conf. on Artificial Intelligence (San Francisco, California, USA February 4–9, 2017)* pp 2576–82
- [17] Samek W, Wiegand T and Klaus-Robert M 2017 Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models *ITU J.: CT Discoveries* **1** 1–10
- [18] Adadi A and Berrada M 2018 Peeking inside the black-box: a survey on explainable artificial intelligence (XAI) *IEEE Access* **6** 52138–60
- [19] Gunning D 2019 DARPA's explainable artificial intelligence (XAI) program *Proc. 24th Int. Conf. on Intelligent User Interfaces, (IUI '19) (New York: ACM)* p ii
- [20] Laura von R, Mayer S, Garcke J, Bauckhage C and Schuecker J 2019 Informed machine learning—towards a taxonomy of explicit integration of knowledge into machine learning (arXiv:1903.12394)
- [21] Schoelkopf B and Smola A 2002 *Learning With Kernels* (Cambridge, MA: MIT Press)
- [22] Camps-Valls G and Bruzzone L (ed) 2009 *Kernel Methods for Remote Sensing Data Analysis* (New York: Wiley)
- [23] Rojo-Álvarez J L, Martínez-Ramón M, Muñoz-Mari J and Camps-Valls G 2018 *Digital Signal Processing With Kernel Methods* (New York: Wiley)
- [24] Berlinet A and Thomas-Agnan C 2011 *Reproducing Kernel Hilbert Spaces in Probability and Statistics* (Boston, MA: Springer)
- [25] Vapnik V N 1998 *Statistical Learning Theory* (New York: Wiley)
- [26] Girosi F, Jones M and Poggio T 1995 Regularization theory and neural network architectures *Neural Comput.* **7** 219–69
- [27] Pérez-Suay A, Laparra V, Mateo-García G, Mu noz-Mari J, Gómez-Chova L and Camps-Valls G 2017 Fair kernel learning *European Conf. on Machine Learning (ECML) (Skopje, Macedonia)*
- [28] Kamishima T, Akaho S and Sakuma J 2011 Fairness-aware learning through regularization approach *2011 IEEE 11th Int. Conf. on Data Mining Workshops* pp 643–50
- [29] Gretton A, Herbrich R, Smola A, Bousquet O and Schölkopf B 2005 Kernel methods for measuring independence *J. Mach. Learn. Res.* **6** 2075–129
- [30] O'Reilly J E and Maritorena S 1997 SeaBAM evaluation data set *The SeaWiFS Bio-Optical Algorithm Mini-Workshop (SeaBAM)* (Santa Barbara, CA: University of California) (available at: <http://seabass.gsfc.nasa.gov/seabam/seabam.html>) (Accessed August 2005)
- [31] O'Reilly J E, Maritorena S, Mitchell B G, Siegel D A, Carder K, Garver S A, Kahru M and McClain C 1998 Ocean color chlorophyll algorithms for SeaWiFS *J. Geophys. Res.* **103** 24937–53
- [32] Verrelst J, Malenovsky Z, Christiaan V der T, Camps-Valls G, Gastellu-Etchegorry J-P, Lewis P, North P and Moreno J 2018 Quantifying vegetation biophysical variables from imaging spectroscopy data: a review on retrieval methods *Surv. Geophys.* **40** 589–629
- [33] Fernández G, Moreno J, Gándia S, Martínez B, Vuolo F and Morales F 2005 Statistical variability of field measurements of biophysical parameters in SPARC-2003 and SPARC-2004 campaigns *Proc. SPARC Workshop*
- [34] Deser C et al 2020 Insights from earth system model initial-condition large ensembles and future prospects *Nat. Clim. Change* **10** 277–86
- [35] Phillips A S, Deser C, Fasullo J, Schneider D P and Simpson I R 2020 Assessing climate variability and change in model large ensembles: a user's guide to the 'climate variability diagnostics package for large ensembles' version 1.0. (10.5065/h7c7-1961)
- [36] Kamishima T, Akaho S, Asoh H and Sakuma J 2012 *Fairness-Aware Classifier With Prejudice Remover Regularizer* (Berlin: Springer) pp 35–50
- [37] Peter B 2020 Rejoinder: invariance, causality and Robustness *Stat. Sci.* **35** 434–6
- [38] Schneider T and Held I M 2001 Discriminants of twentieth-century changes in earth surface temperatures *J. Clim.* **14** 249–54
- [39] Wills R C J, Battisti D S, Armour K C, Schneider T and Deser C 2020 Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations *J. Clim.* **33** 8693–719
- [40] Hasselmann K 1993 Optimal fingerprints for the detection of time-dependent climate change *J. Clim.* **6** 1957–71
- [41] Hegerl G and Zwiers F 2011 Use of models in detection and attribution of climate change *WIREs Clim. Change* **2** 570–91
- [42] Santer B D et al 2019 Celebrating the anniversary of three key events in climate change science *Nat. Clim. Change* **9** 180–2
- [43] Barnes J D, Balaguer L, Manrique E, Elvira S and Davison A W 1992 A reappraisal of the use of DMSO for the extraction and determination of chlorophylls *a* and *b* in lichens and higher plants *Environ. Exp. Bot.* **32** 85–100
- [44] Sippel S, Meinshausen N, Fischer E M, Székely Eo and Knutti R 2020 Climate change now detectable from any single day of weather at global scale *Nat. Clim. Change* **10** 35–41
- [45] Madakumbura G D, Thackeray C W, Norris J, Goldenson N and Hall A 2021 Anthropogenic influence on extreme precipitation over global land areas seen in multiple observational datasets *Nat. Commun.* **12** 3944
- [46] Barnes E A, Hurrell J W, Ebert-Uphoff I, Anderson C and Anderson D 2019 Viewing forced climate patterns through an ai lens *Geophys. Res. Lett.* **46** 13389–98
- [47] Sippel S, Meinshausen N, Székely E, Fischer E, Pendergrass A G, Lehner F and Knutti R 2021 Robust detection of forced warming in the presence of potentially large climate variability *Sci. Adv.* **7** 43

- [48] Meinshausen N 2018 Causality from a distributional robustness point of view *2018 IEEE Data Science Workshop (DSW)* pp 6–10
- [49] Pan S J and Yang Q 2009 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
- [50] Olonscheck D, Rugenstein M and Marotzke J 2020 Broad consistency between observed and simulated trends in sea surface temperature patterns *Geophys. Res. Lett.* **47** e2019GL086773
- [51] Fredriksen H-B, Berner J, Subramanian A C and Capotondi A 2020 How does el niño-southern oscillation change under global warming—a first look at CMIP6 *Geophys. Res. Lett.* **47** e2020GL090640
- [52] Seager R, Cane M, Henderson N, Lee D-E, Abernathey R and Zhang H 2019 Strengthening tropical pacific zonal sea surface temperature gradient consistent with rising greenhouse gases *Nat. Clim. Change* **9** 517–22
- [53] Raissi M, Perdikaris P and Karniadakis G E 2017 Machine learning of linear differential equations using gaussian processes *J. Comput. Phys.* **348** 683–93
- [54] Cortés-Andrés J and Camps-Valls G (PKL) Physics-aware Kernel Learning 2022 (available at: <https://github.com/IPL-UV/PKL>)
- [55] Kimeldorf G S and Wahba G 1970 A correspondence between Bayesian estimation on stochastic processes and smoothing by splines *Ann. Math. Stat.* **41** 495–502
- [56] Fukumizu K, Gretton A, Sun X and Schölkopf P B 2008 Kernel measures of conditional dependence *Advances in Neural Information Processing Systems* vol 20, ed J C Platt, D Koller, Y Singer and S T Roweis (Curran Associates, Inc.) pp 489–96
- [57] Schölkopf B and Smola A 2002 *Learning with Kernels—Support Vector Machines, Regularization, Optimization and Beyond* (Cambridge, MA: MIT Press)