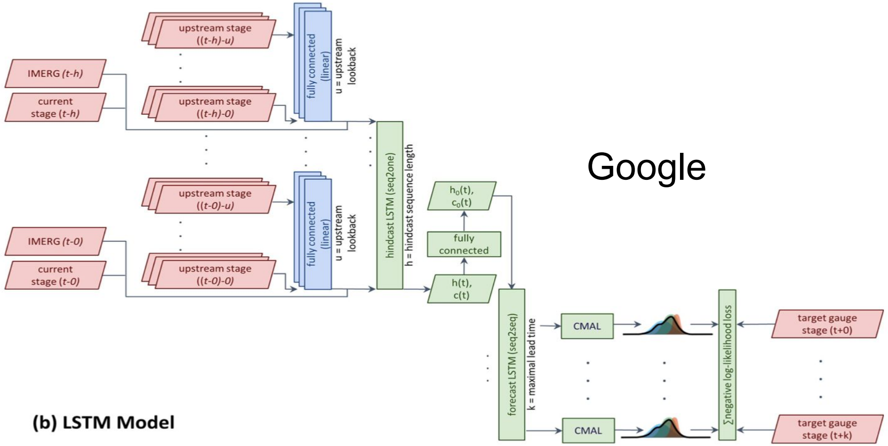
- Google's forecasting system consists of four subsystems: data validation, stage forecasting, inundation modeling, and alert distribution; stage forecasting is modeled with the long short-term memory (LSTM) networks and the linear models

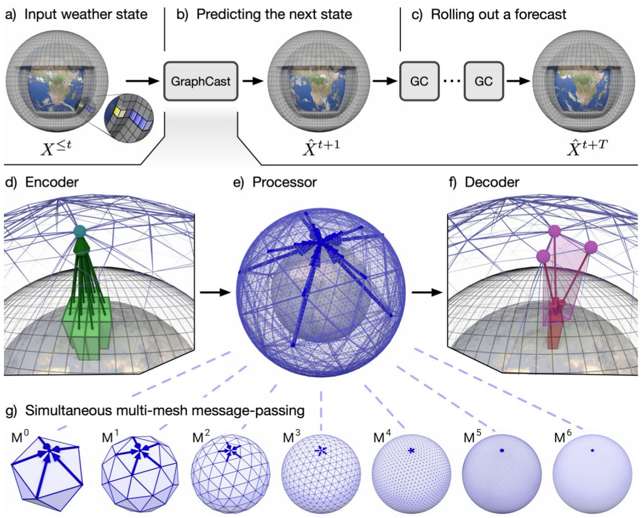
- GraphCast capitalises on GNN’s ability to model arbitrary sparse interactions by introducing internal multi-mesh representation, which has homogeneous spatial resolution over the globe, and allows long-range interactions within few message-passing steps

- In certain domains, incorporating domain-specific knowledge into the input features can improve performance. For example, in computer vision tasks, handcrafted features like edges or texture descriptors can provide additional information to the neural network



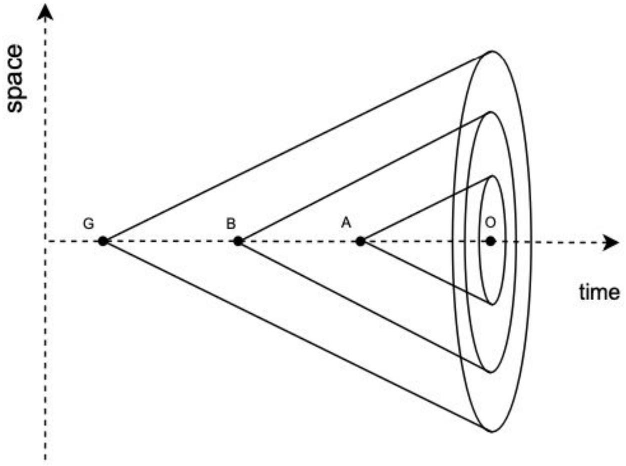
(b) LSTM Model

Google



GraphCast

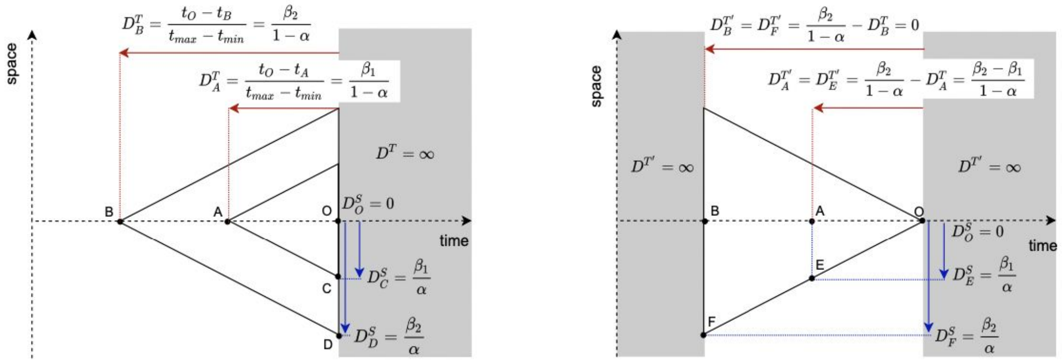
- Most predictive spatio-temporal machine learning models transform the original problem into a multiple regression task, where the target variable is the future value of the series and the predictors are previous past values of the series up to a certain time window
- Assumes that the future conditions depend on recently observed conditions in the same location
- Certain spatio-temporal variables of interest depend not only on the recent past conditions at the same location, but also on recent past conditions of nearby locations
- The “spatial radius of influence” may decrease for older observations such that observations that are more proximate in both time and location have a larger influence on the outcome variable than temporally and spatially distance observations.



$$D_A^T = \begin{cases} \frac{t_O - t_A}{t_{max} - t_{min}}, & \text{if } t_A < t_O \\ \infty, & \text{otherwise} \end{cases}$$

$$D_A = \alpha \cdot D_A^S + (1-\alpha) \cdot D_A^T \leq \beta$$

$$\mathcal{N}_o^\beta = \{k \in \mathcal{D} : D_{o,k} < \beta\}$$



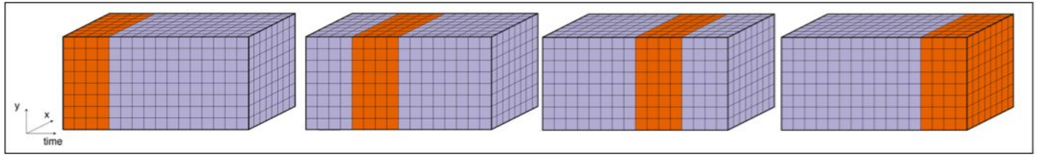
(a) Original cone

(b) Reversed cone

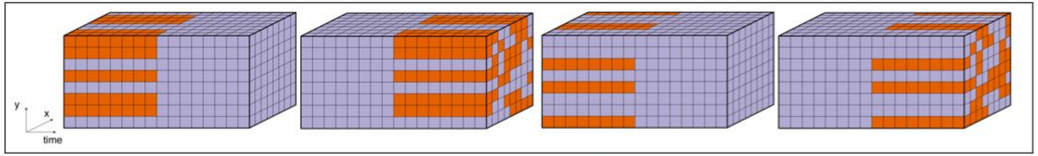
- Commonly used performance estimation procedures such as cross-validation (CV) and out-of-sample (OOS) validation face challenges due to the implicit dependence between observations in spatiotemporal datasets

- Standard cross-validation leads to over-optimistic estimates in spatio-temporal settings

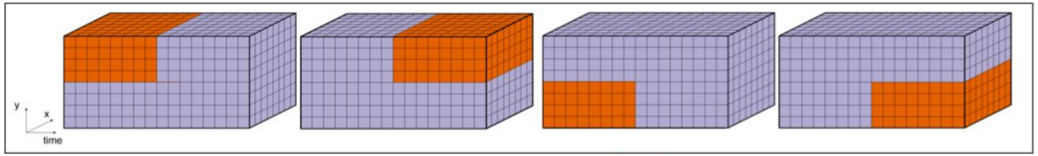
- Blocking data in space and/or in time is useful in mitigating CV’s bias to underestimate error



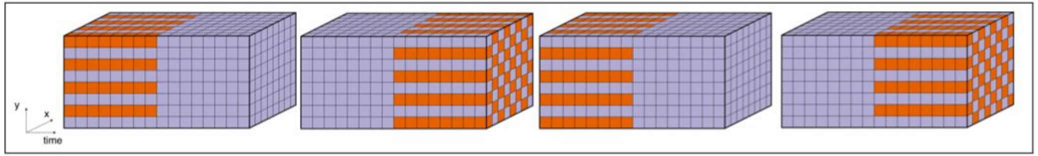
(a) Time block CV (CV-Tb)



(b) Spatio-temporal block CV (CV-STb)



(c) Spatio-temporal contiguous block CV(CV-STb-cont)



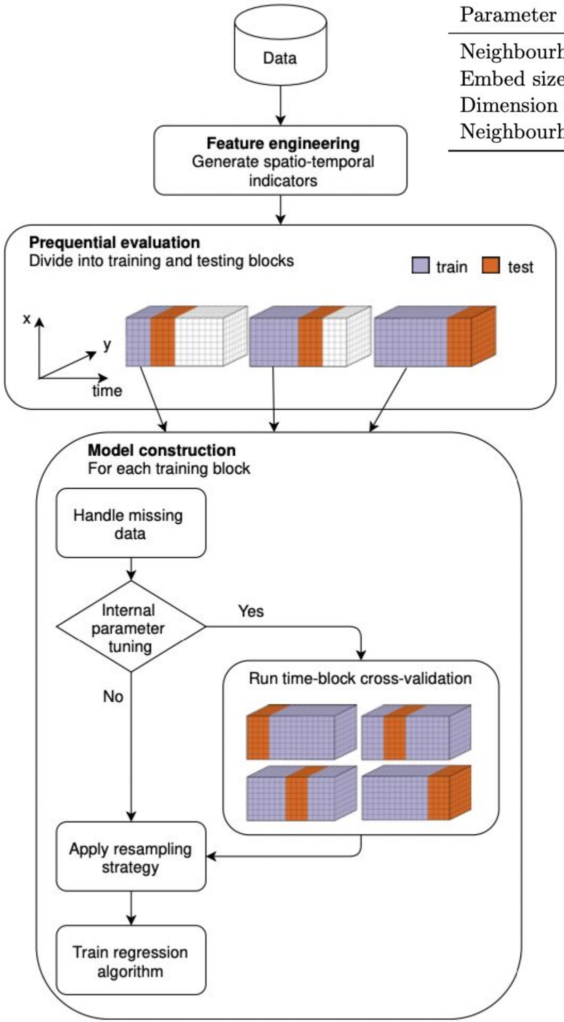
(d) Spatio-temporal systematic block CV (CV-STb-sys)

Figure 3.3: Block cross-validation methods that have prequential equivalents. Folds used for training in lighter lilac; folds used for testing in dark orange. Time flows left to right

- Identify methods for optimizing parameters of spatio-temporal cone shape, α and β, as well as orientation, determined by D_a^T, for a given data set (e.g. producing a grid search algorithm or using numerical optimization, where possible)

- Apply block cross-validation to evaluate parameter tuning methods on a variety of spatio-temporal data sets using different modeling approaches (linear regression, SVM, regression trees, random forests, CNN, LSTM, GNN)

- If sufficient time remains, develop and publish a package in R to select optimal spatio-temporal neighbourhoods for different spatial data types (raster, point, polygon, and line)



Parameter	Search space
Neighbourhood type	cone, reversed
Embed size (K)	4, 8
Dimension weight α	0.1, 0.25, 0.5, 0.75, 0.9
Neighbourhood radii β	$\{0.01, 0.02, 0.03\}$, $\{0.02, 0.03, 0.04\}$

- So far have gathered list of key terms (mostly for my own reference) and summarised 43 relevant papers in literature review

Area	Paper	Authors	Year	Publication	Summary
Hyrdology	**A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources**	Tyralis, Papacharalampous, & Langousis	2019	*Water*	– Review random forest applications in water resources, highlight the potential of the original algorithm and its variants, and assess the degree of RF exploitation in a diverse range of applications – Random forests are said to belong to the class of data-driven models – Provides literature review of various statistical models applied in hyrdology
Spatio-temporal forecasting	**A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates**	Lindström et al.	2014	*Environmental and Ecological Statistics*	– Extend work on air pollution forecasting to include spatio-temporal covariates, e.g. output of deterministic point-prediction air pollution model – The authors' proposed method can also include geographic covariates, and it was implemented in an R package, SpatioTemporal (removed) – The model uses temporal basis functions to account for the temporal variability in data; to account for spatial variability in the temporal structure, the basis functions are modulated by spatially varying coefficients; the coefficients are modeled using universal kriging, where the linear trend contains Geographic Information System (GIS) covariates – The parameters of the model consist of regression parameters for the spatio- temporal and geographic covariates – The spatially varying coefficients for the temporal functions are treated as spatial fields with a universal kriging structure, allowing the temporal structure to vary between locations – The residual space-time field is assumed to be independent in time with stationary, parametric spatial covariance – https://sites.stat.washington.edu/people/peter/591/labs/Lab4/Lab4.SpatioTemporal.html
Spatio-temporal forecasting	**A novel framework for spatio-temporal prediction of environmental data using deep learning**	Amato et al.	2020	*Nature Scientific Reports*	– While Deep Learning models have proved to be able to capture spatial, temporal, and spatio-temporal dependencies through their automatic feature representation learning, the problem of the interpolation of continuous spatio-temporal fields measured on a set of irregular points in space is still under-investigated – Introduce a framework for spatio-temporal prediction of climate and environmental data using deep learning by showing how spatio-temporal processes can be decomposed in terms of a sum of products of temporally referenced basis functions, and of stochastic spatial coefficients which can be spatially modelled and mapped on a regular grid, allowing the reconstruction of the complete spatio-temporal signal – No ML-based methodology has been proposed to solve spatio-temporal interpolation with spatially irregular ground measurements as input – Propose an advanced methodological framework to reconstruct a spatio-temporal field on a regular grid using spatially irregularly distributed time series data – It is difficult to realistically reproduce the spatial, temporal and spatio-temporal dependencies in the data; one way of keeping into account these dependencies is to adopt a basis function representation; numerous basis functions can be used to decompose spatio-temporal data, including complete global functions as with Fourier analysis, and non-orthogonal bases as with wavelets. Here we use the reduced-rank basis obtained through a principal component analysis (PCA) – While in ML PCA is generally applied to reduce the dimensionality of the input space, here we use it on the output space—i.e. the spatio-temporal target variable we want to interpolate—in order to decompose the data into fixed temporal bases and their corresponding spatial coefficients; the latter are then considered as the target variable in a spatial regression problem, which can be solved using any ML technique – First, a basis function representation is used to extract fixed temporal bases from the spatio-temporal observations; then, the stochastic spatial coefficients corresponding to each basis functions are modelled jointly at any desired spatial location with a DL regression technique; finally, the spatio-temporal signal is recomposed, returning a spatio-temporal interpolation of the field – The proposed methodology was applied on two different datasets: a simulated spatio-temporal field, and a real-world dataset of temperature
Spatio-temporal forecasting	**A spatio-temporal autoregressive model for monitoring and predicting COVID infection rates**	Congdon	2022	*Journal of Geographical Systems*	– Consider different specifica- tions of autoregressive dependence in incidence counts as these may considerably impact on adaptivity in epidemic situations; introduce parameters to allow temporal adaptivity in autoregressive dependence – The model allows for both spatial and time variation in autoregressive effects; authors assess sensitivity in short-term pre- dictions and fit to specification (spatial vs space-time autoregression, linear vs log- linear, and form of space decay), and show improved one-step ahead and in-sample prediction using space-time autoregression including temporal adaptivity – The general form of this representation is termed the autoregressive conditional Poisson (ACP); to generalize these representations to area-time infection count data, one may add lags to infection counts in spatially close areas
Spatio-temporal forecasting	**An autoregressive spatio-temporal precipitation model**	Sigrist, Künsch, & Stahel	2011	*Procedia Environmental Sciences*	– Assumes that precipitation follows a censored and power-transformed normal distribution; precipitation is linked to covariates via regression – Spatial and temporal dependencies are accounted for by a latent Gaussian variable that follows a Markovian temporal evolution combined with spatially correlated innovations; such a specification allows for nonseparable covariances in space and time
Spatio-temporal forecasting	**An efficient implementation for spatial–temporal Gaussian process regression and its applications**	Zhang et al.	2023	*Automatica*	– Spatial–temporal Gaussian process regression is a popular method for spatial–temporal data modeling – By exploring the Kronecker structure of the state-space model realization of the spatial–temporal Gaussian process, it is possible to further reduce the computational complexity
Feature selection	**An Introduction to Variable and Feature Selection**	Guyon & Elisseeff	2003	*Journal of Machine Learning Research*	– Provide a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods
Spatio-temporal forecasting	**Analyzing Big Spatial and Big Spatiotemporal Data: A Case Study of Methods and Applications**	Chandola et al.	2015	*Handbook of Statistics*	– Study approaches to handle big spatial and spatiotemporal data by closely looking at the computational and I/O requirements of several analysis algorithms for such data

Area	Concept	Acronym	Defiition	Equation
Spatial autocorrelation	**Range**	-	A characteristic of spatial dependence that represents the distance beyond which the spatial correlation is considered to be negligible; the maximum spatial extent over which nearby locations exhibit significant similarity	-
Spatial autocorrelation	**Sill**	-	A characteristic of spatial dependence that represents the maximum **semivariance** value observed in a **variogram plot**; represents the variance of data when the lag distance approaches infinity; provides an upper bound for the **ser**	-
Spatial autocorrelation	**Nugget**	-	A characteristic of spatial dependence that represents the **semivariance** value at a lag distance of zero or very close to zero; represents the spatial variance at very short distances; associated with measurement errors that are not	-
Spatial autocorrelation	**Spherical variogram model**	-	A model used to describe the spatial correlation structure for a **variogram** that assumes spatial correlation reaches a maximum value within a circular **range** then levels off beyond that **range**; increases in **semivariance** until the **sill**	-
Spatial autocorrelation	**Tetraspherical variogram model**	-	A model used to describe the spatial correlation structure for a **variogram** that extends the **spherical variogram model** by allowing for a slower decrease in **semivariance** beyond the **range**	-
Spatial autocorrelation	**Pentaspherical variogram model**	-	A model used to describe the spatial correlation structure for a **variogram** that extends the **tetraspherical variogram model** by allowing for additional parameters to capture more gradual transitions in **semivariance**	-
Spatial autocorrelation	**Circular variogram model**	-	A model used to describe the spatial correlation structure for a **variogram** that assumes spatial correlation is constant within a circular **range** then drops off sharply outside that **range**; the simplest **variogram** model	-
Spatial autocorrelation	**Moran's *I***	-	A measure of spatial autocorrelation where negative one indicates perfectly negative correlation, zero indicates no correlation, and positive one indicates perfect correlation	$I = N\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})/W\sum\{i=1\}^{n}(x_i - \bar{x})^2$
Spatial in-filling	**Leaves**	-	The areas or regions in a image that are missing and need to be filled in	-
Spatio-temporal forecastin	**Spatial stationarity**	-	When the spatial correlation structure of spatial data remains constant over time	-
Spatio-temporal forecastin	**Kriging**	-	A spatial interpolation technique used to estimate values at unobserved locations based on known values at nearby locations using a fitted **variogram** model; finds the optimal linear combination of nearby known values; assumes **s**	-
Spatio-temporal forecastin	**Co-kriging**	-	An extension to **kriging** that allows for the use of predictor variables beyond the target variable	
Spatio-temporal forecastin	**Variogram**	-	A measure of the spatial dependence or auto-correlation of a spatial dataset; describes how the similarity of values between pairs of spatial locations changes with distance; used in **kriging**; also called a semivariogram or an experir	-
Spatio-temporal forecastin	**Semivariance**	-	A measure of the average difference between pairs of spatial locations within each lag distance class; quantifies differences as a function of the lag distance; used in calculating **variograms**	-
Spatio-temporal forecastin	**Variogram Plot**	-	A plot of the **semivariance** values against the lag distances on a scattler plot; visualises the relationship between spatial separation and **semivariance**; a **variogram** describes this function	-
Spatio-temporal forecastin	**Spatio-temporal autoregressive mod**	STAR	A spatio-temporal model that explicitly models spatio-temporal dependence by considering the influence of neighbourhing locations at previous time points; essentially extends spatio autoregressive models to also include temporal	-
Spatio-temporal forecastin	**Dynamic spatio-temporal model**	-	A spatio-temporal model that incorporates dynamic components that account for changes in spatial dependencies and temporal trends over time; includes **state-space models** and Bayesian hierarchical models	-
Spatio-temporal forecastin	**Imbalanced task**	-	Wwhen certain ranges of values in the target variable are more important to the end-user, but severely under-represented in the training data such as predicting extreme values	-
Spatio-temporal forecastin	**Block Cross-Validation**	CV-b	A method similar to K-fold cross-validation where each fold is a sequential, non-interrupted time series instead of each fold containing a random subset of observations	-
Spatio-temporal forecastin	**Spatio-temporal block cross-validatio**	CV-STb	A form of **block cross-validation** where time is grouped in sequential blocks and locations are simultaneously grouped together randomly, in contiguous geographic regions, or in a systematic checkered pattern	
Spatio-temporal indicators	**Time-varying spatial weights**	-	A set of spatial weights stored in matrix format that vary over time; reflect changes in spatial dependence over time	-
Spatio-temporal indicators	**Spatio-temporal distance**	-	The distance between any two points in the space-time dimension	$D_{u,v} = \alpha \cdot d_{i,j} + (1 - \alpha) \cdot t_{k,l}$ for z_i^k and z_k^l
Spatio-temporal indicators	**Great-circle distance**	-	The shortest distance between two points on the surface of a sphere, measured along the surface of the sphere; can be used to describe the spatial distance component of the **spatio-temporal distance**	-
Spatio-temporal indicators	**Spatio-temporal neighbourhood**	-	The set of points within a certain **spatio-temporal distance** of a given point	$\mathcal{N}_o\text{^}\beta = \{k \in \mathcal{D}: \mathcal{D}_{o,k} < \beta\}$
Spatio-temporal indicators	**Spatio-temporal weight**	-	The inverse of the **spatio-temporal distance** between two points	$w_{o,x} = 1/D_{o,x}$
Spatio-temporal indicators	**Spatio-temporal typical value**	-	A spatio-temporal indicator equal to the average value of target variable within the **spatio-temporal neighbourhood** of the target location; weighted version multiplies the **spatio-temporal weight** to the data points	$\bar{w}(\mathcal{N}_o\text{^}\beta) = 1/\lvert\mathcal{N}\rvert \cdot \sum(x \in \mathcal{N}_o\text{^}\beta)\, x$
Spatio-temporal indicators	**Spatio-temporal spread**	-	A spatio-temporal indicator equal to the standard deviation of the target variable within the **spatio-temporal neighbourhood** of the target location; weighted version uses the weighted **spatio-temporal typical value**	$\sigma(\mathcal{N}_o\text{^}\beta) = \surd[1/\lvert\mathcal{N}\rvert \cdot \sum(x \in \mathcal{N}_o\text{^}\beta)(x - \bar{w}(\mathcal{N}_o\text{^}\beta))^2]$
Spatio-temporal indicators	**Spatio-temporal tendency**	-	A spatio-temporal indicator equal to the ratio between two **spatio-temporal weights** calculated with different **spatio-temporal neighbourhood** sizes; weighted version uses the weighted **spatio-temporal typical value**	$W\text{^}\{\beta_1,\beta_2\}(\mathcal{N}_o\text{^}\beta) = \bar{w}(\mathcal{N}_o\text{^}\beta_1)/\bar{w}(\mathcal{N}_o\text{^}\beta_2)$
Temporal autocorrelation	**State-space model**	-	A statistical model that describes the evolution of a system of time; the current state typically is represented as a linear combination of the previous state and random error; also called a dynamic linear model or a hidden Markov r	-
Time series forecasting	**Time-delay embedding**	-	A process of forecasting future values by including past values and predictors as lagged variables in a prediction model	-
Time series forecasting	**Technical indicators**	-	The statistical information that is used to expand the set of lagged variables used in **time-delay embedding**; statistical summaries of certain properties of the time series; both distance and time components are normalised	-
Time series forecasting	**Sliding Window Approach**	-	An approach to time series forecasting that uses predictions of previous time periods as predictors for future time periods; can exponentially increase errors	-

- Next step is data collection, methods that are used are highly dependent on the type of data that is available; more detailed with high space/time resolution is preferable

- Domain expertise required

scientific **data**

Check for updates

OPEN

DATA DESCRIPTOR

Caravan - A global community dataset for large-sample hydrology

Frederik Kratzert[1 ✉], Grey Nearing[2], Nans Addor[3,4], Tyler Erickson[5], Martin Gauch[6], Oren Gilon[7], Lukas Gudmundsson[8], Avinatan Hassidim[7], Daniel Klotz[6], Sella Nevo[7], Guy Shalev[7] & Yossi Matias[7]

High-quality datasets are essential to support hydrological science and modeling. Several CAMELS (Catchment Attributes and Meteorology for Large-sample Studies) datasets exist for specific countries or regions, however these datasets lack standardization, which makes global studies difficult. This paper introduces a dataset called *Caravan* (a series of CAMELS) that standardizes and aggregates seven existing large-sample hydrology datasets. Caravan includes meteorological forcing data, streamflow data, and static catchment attributes (e.g., geophysical, sociological, climatological) for 6830 catchments. Most importantly, Caravan is both a dataset and open-source software that allows members of the hydrology community to extend the dataset to new locations by extracting forcing data and catchment attributes in the cloud. Our vision is for Caravan to democratize the creation and use of globally-standardized large-sample hydrology datasets. Caravan is a truly global open-source community resource.

THE GLOBAL SYNTHETIC DATASET

News: *In December 2022, IOM released The Global Victim-Perpetrator Synthetic Dataset produced using an updated version of Synthetic Data Showcase with added support for differential privacy. The resulting dataset describes victim-perpetrator relations. It is CTDC's second synthetic dataset, and the first to provide the guarantee of differential privacy.*

Microsoft Research has worked with IOM to develop a new algorithm to derive "synthetic data" from CTDC's sensitive victim case data. Rather than systematically redacting cases, which results in a substantial amount of data being suppressed, the algorithm generates a synthetic dataset that accurately preserves the statistical properties and relationships in the original data. Representative data on all of CTDC's victim of trafficking cases are now available as a downloadable data file thanks to the new algorithm.