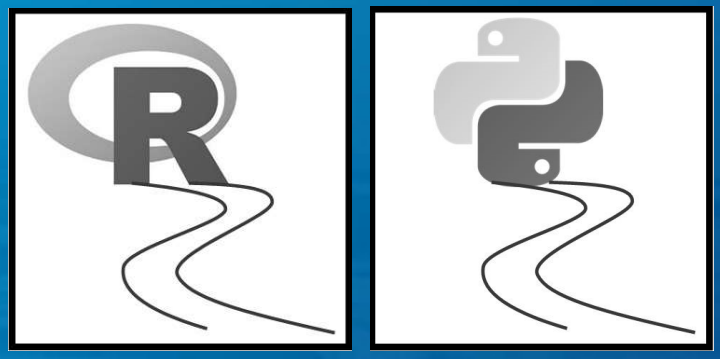
#R #Python #hydrology #timeseries

Lima, 6-nov 2023

VI CONGRESO NACIONAL DEL AGUA Perú 2023

Ciencia de datos en Hidrología
Data Science in Hydrology

Pedro Rau, PhD
Hidrólogo



https://github.com/hydrocodes

VI CONA PERÚ 2023

UTEC UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

cita CENTRO DE INVESTIGACIÓN Y TECNOLOGÍA DEL AGUA · UTEC

UTEC Posgrado

seguridad hídrica y cambio climático
grupo de investigación

- ✓ **Instructor:**

  Pedro Rau, PhD

  Profesor asociado Universidad de Ingeniería y Tecnología UTEC – Dpto. Ing. Civil y Ambiental. Investigador principal Centro de Investigación y Tecnología del Agua CITA. Hidrólogo investigador en redes internacionales: FR, EC, US, UK … Consultor en Hidrología y Recursos Hídricos.
- ✓ **E.mail:** pedro.rau.ing@gmail.com
- ✓ **Sitio web:** http://pedrorau.blogspot.com
- ✓ **Repositorio del curso:** https://github.com/hydrocodes/datascience
- ✓ **Requisitos:** Conocimiento básicos en Hidrología y Estadística

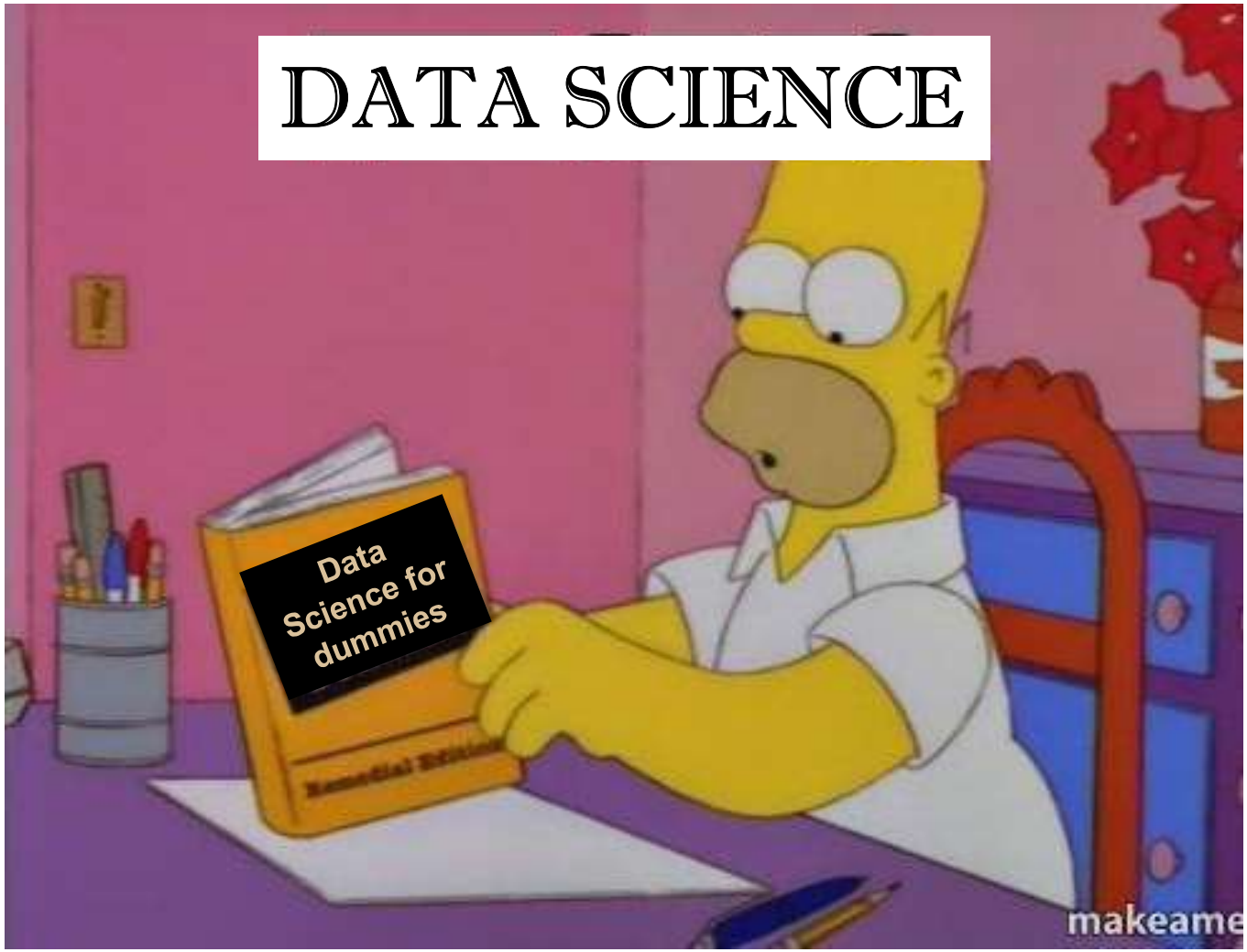
Contenido del taller

1. Introducción a la ciencia de datos
2. Series de tiempo
3. Entornos R y Python

Metodología « Hands-on »

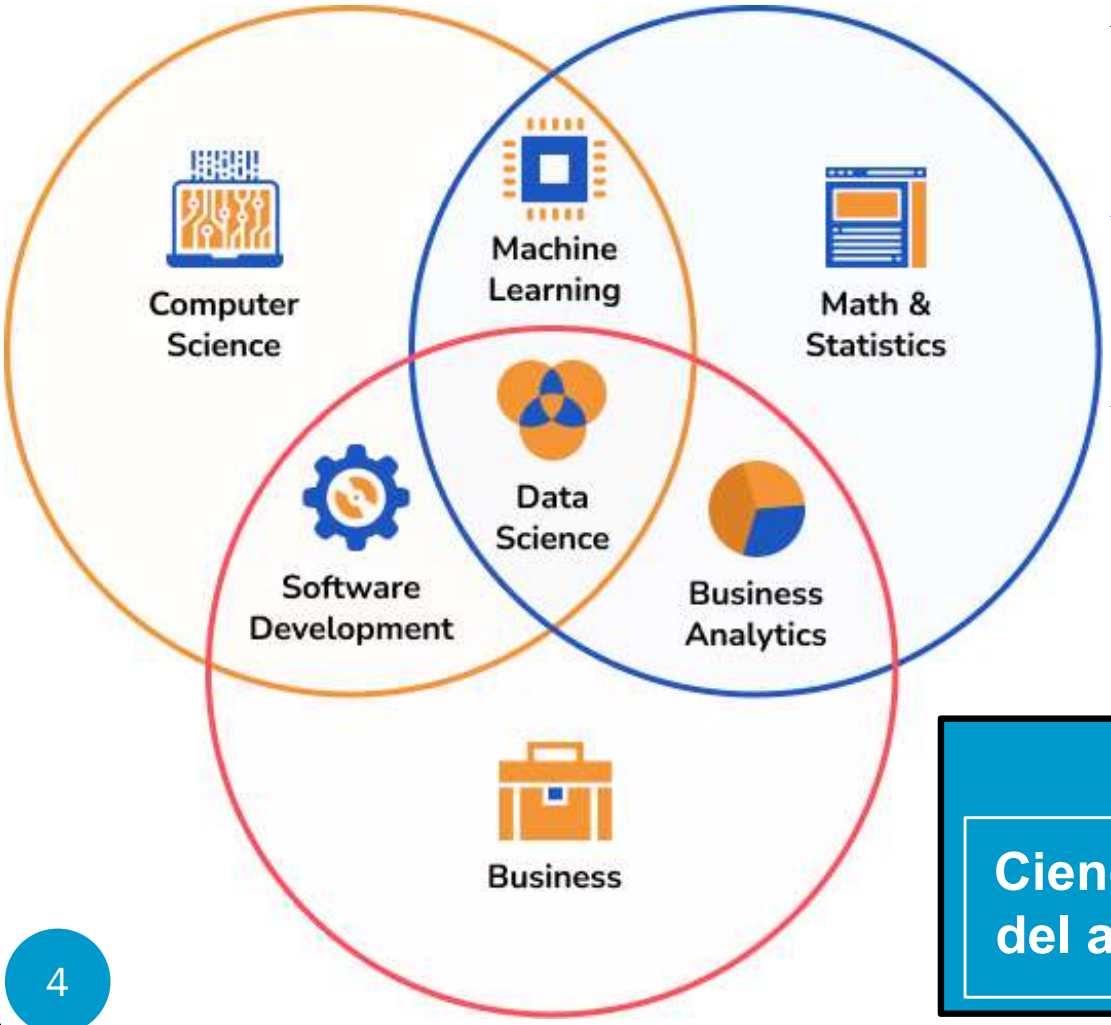
- Lecturas base y planteamiento de ejercicios
- Creación e interpretación de ficheros *.R con comentarios
- Resolución de bugs

1. Introducción a la ciencia de datos en hidrología



1.1 Ciencia de datos

Combina matemáticas y estadística, programación especializada, análisis avanzado, inteligencia artificial y aprendizaje automático con experiencia en el ciclo hidrológico para **descubrir conocimientos** prácticos ocultos en los datos de una cuenca o unidad de gestión. Estos conocimientos se pueden utilizar para guiar la toma de decisiones y la planificación estratégica.

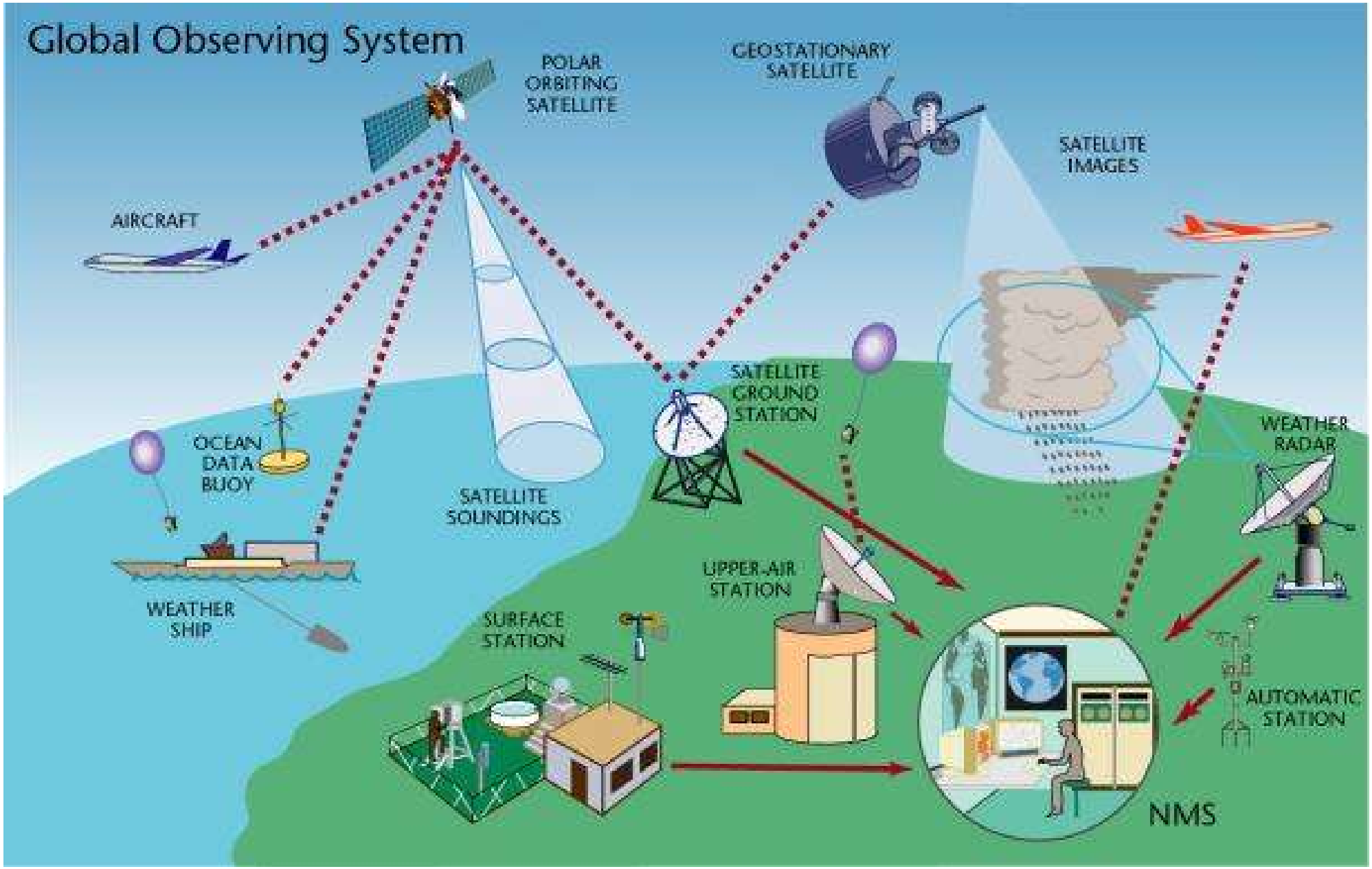


- ✓ Una persona genera **1.7 MB** de datos por segundo.
- ✓ **2.5 x 10^15 MB** de datos por día en la web.
- ✓ **1 millón de datos**: Registros cada 15 min a lo largo de un año en el datalogger de un pluviómetro.

Hidroinformática

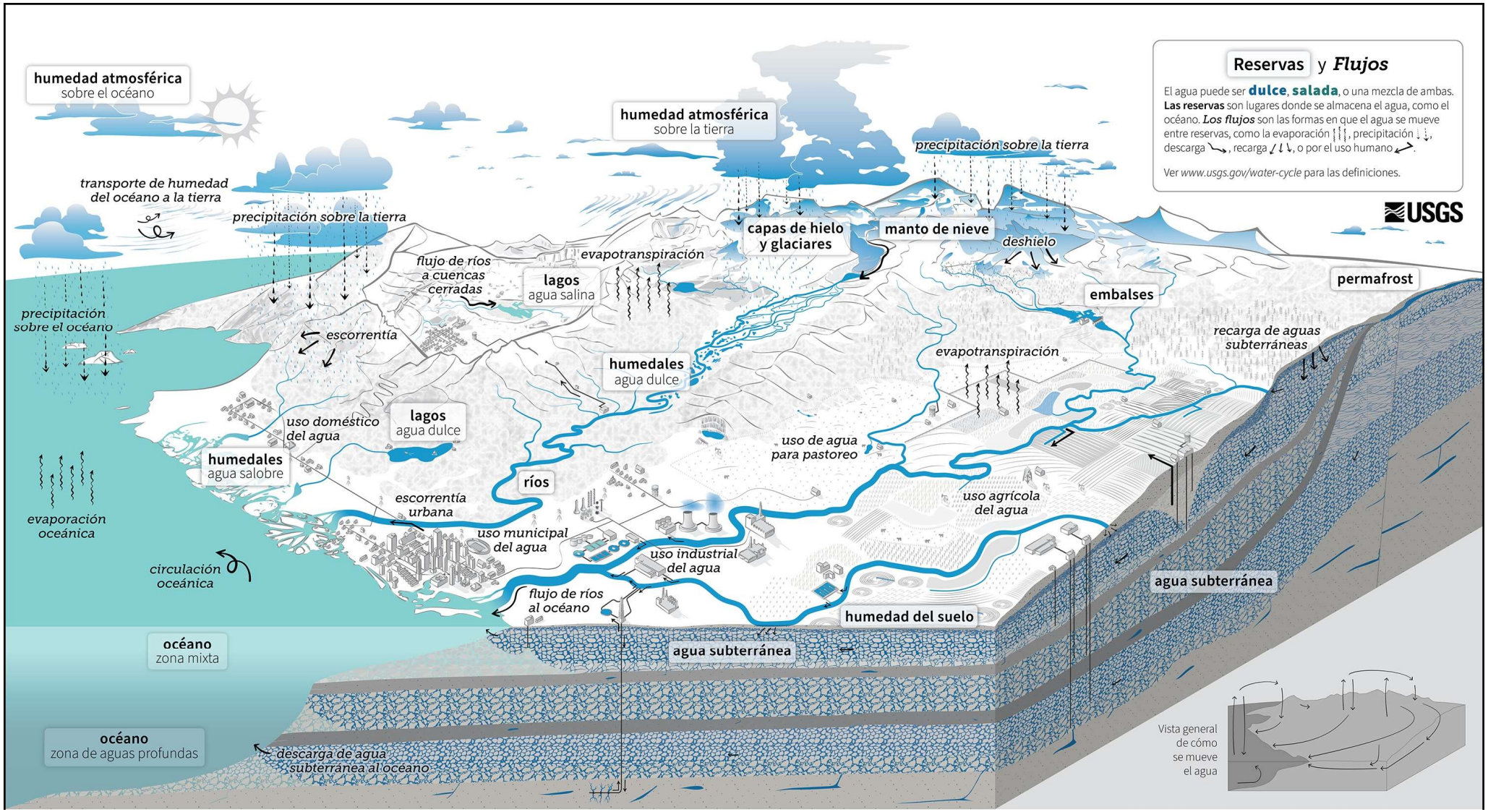
Ciencias del agua	Ciencia de datos	Ciencias de la computación

1.2 Sistema de observación global



WMO

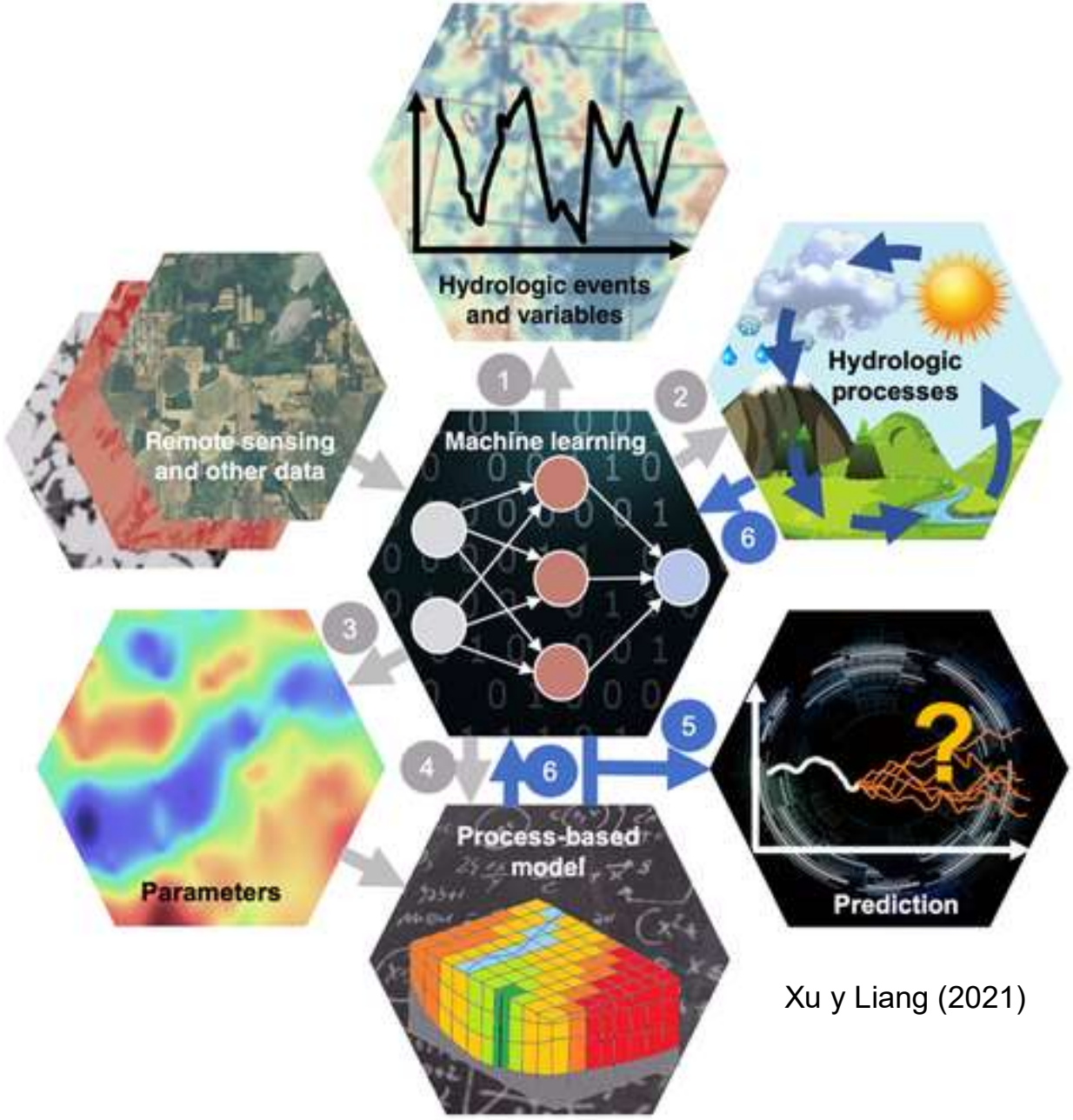
1.3 El "nuevo" ciclo hidrológico



USGS (2022)

Dinámico, no estacionario, inherente a la sociedad, huella hídrica

1.4 Machine learning, Deep learning e Inteligencia Artificial

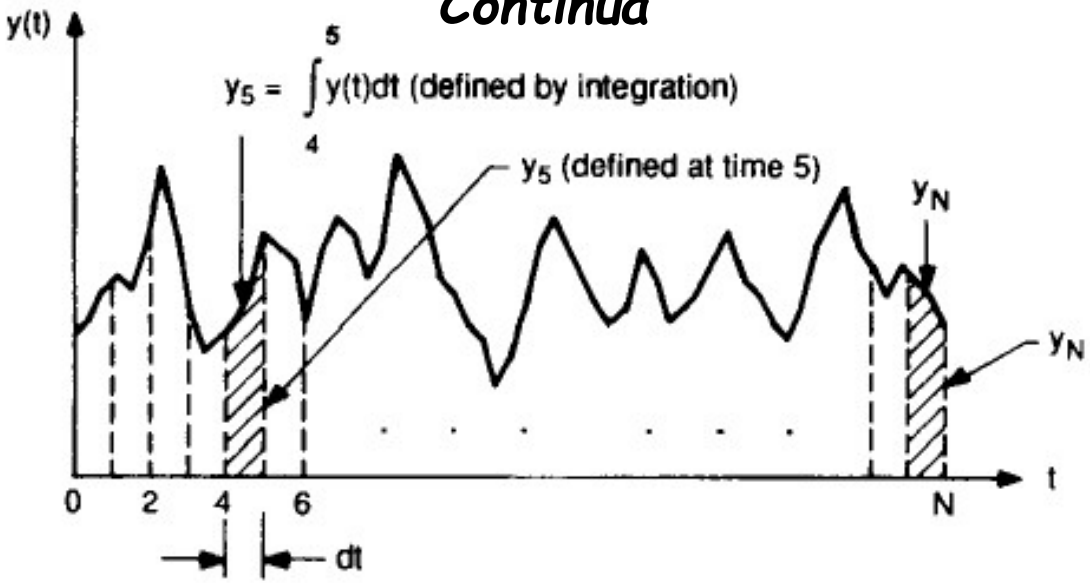


Xu y Liang (2021)

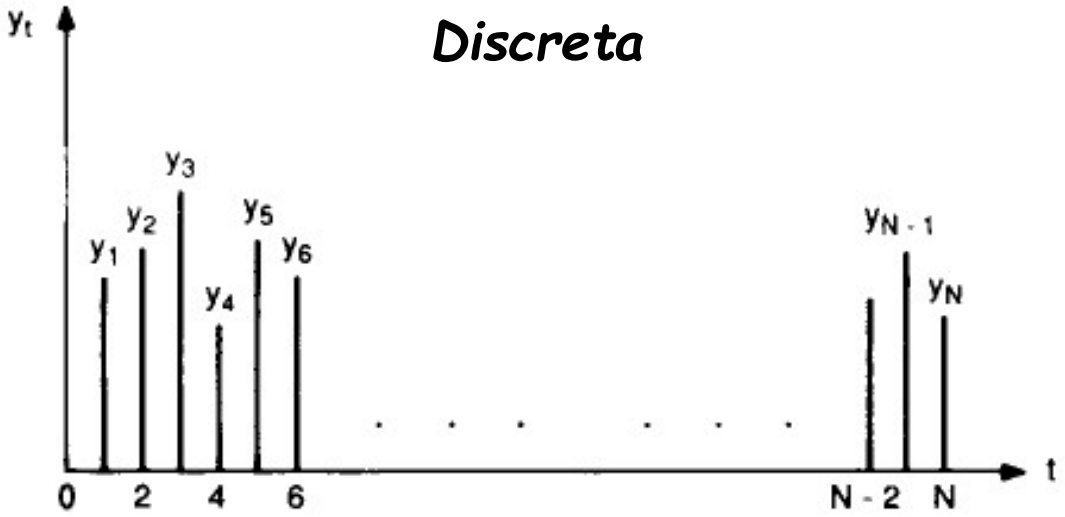
2. Tratamiento de datos hidrológicos

2.1 Tipos o categorías de series de tiempo

Continua



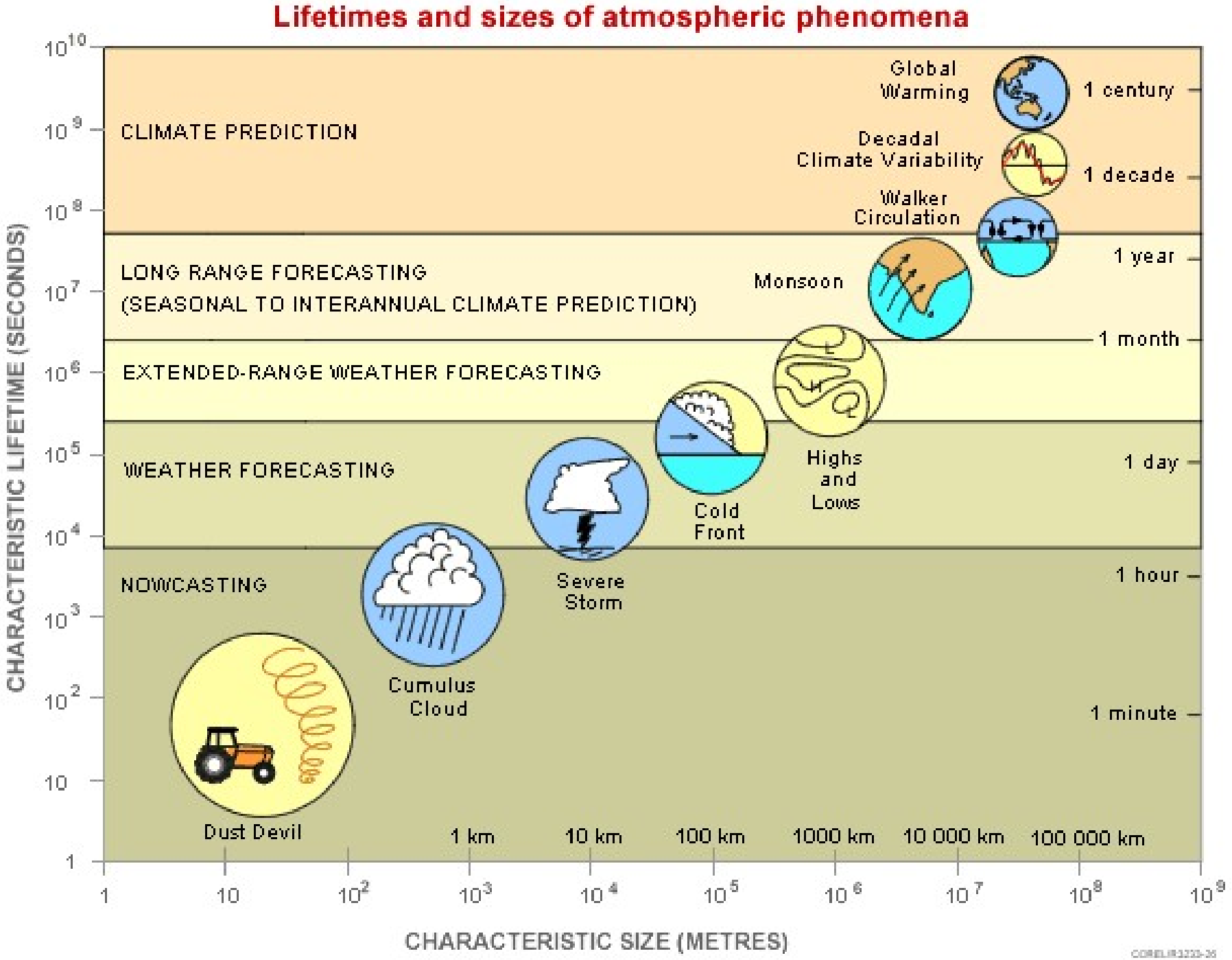
Discreta



- ✓ Unicas (univariado)
- ✓ Multiples (multivariado)
- ✓ Independientes, no correlacionadas; autocorrelacionados o dependientes
- ✓ Intermitentes *(e.g. con ceros)*
- ✓ De conteo *(e.g. dias lluviosos)*
- ✓ Regulares o irregulares en espacios de tiempo
- ✓ Estacionarias (sin tendencia, ni salto ni ciclicidad / periodicidad) y no-estacionarias

Salas (1996)

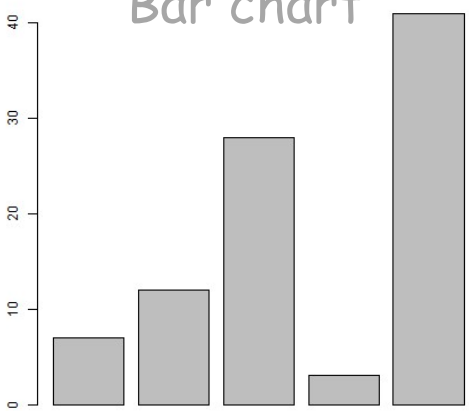
2.2 Escalas de tiempo espacio-temporales



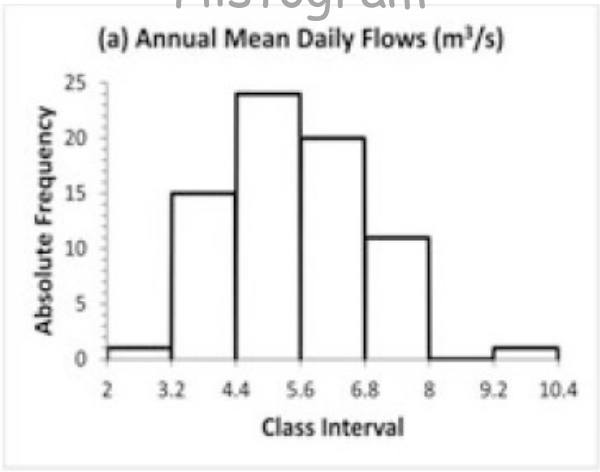
2.3 Análisis preliminar de datos hidrológicos

a. Algunos tipos de representaciones gráficas

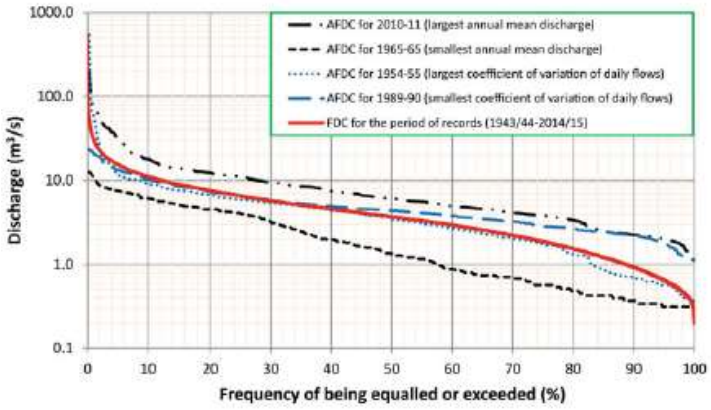
Diagrama de barras
Bar chart



Histograma
Histogram



Curvas de duración
Duration curves



b. Estadísticos descriptivos

Parámetro de la población	Estadística de la muestra
1. *Punto medio*	
Media aritmética	
$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Mediana	
x tal que $F(x) = 0.5$	Valor de la información en el 50o. percentil
Media geométrica	
antilog $[E(\log x)]$	$\left(\prod_{i=1}^{n} x_i\right)^{1/n}$
2. *Variabilidad*	
Varianza	
$\sigma^2 = E[(x-\mu)^2]$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
Desviación estándar	
$\sigma = \{E[(x-\mu)^2]\}^{1/2}$	$s = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}$

Coeficiente de variación

$CV = \frac{\sigma}{\mu}$ $\qquad CV = \frac{s}{\bar{x}}$

3. *Simetría*

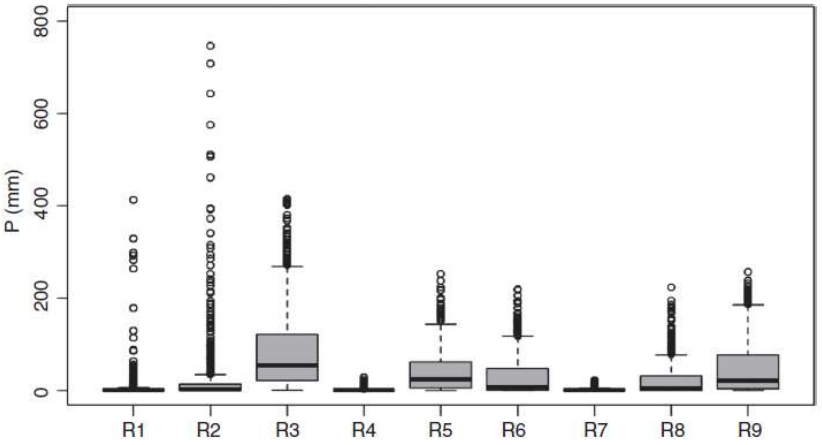
Coeficiente de asimetría (oblicuidad)

$\gamma = \frac{E[(x-\mu)^3]}{\sigma^3}$ $\qquad C_s = \frac{n\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n-1)(n-2)s^3}$

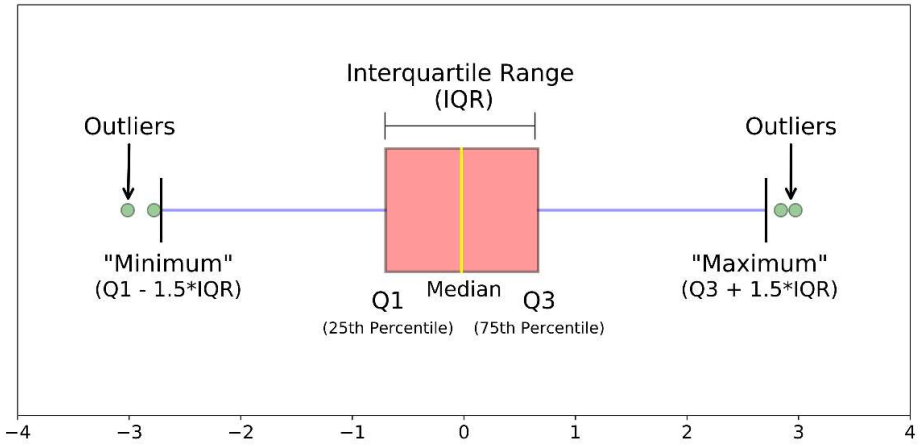
Chow (1994)
Naghettini (2017)

c. Algunos métodos exploratorios

Diagrama de cajas
Box plot



Rau et al (2017)



Naghettini (2017)

d. Asociación de datos

Gráfico de dispersion
Scatter plot

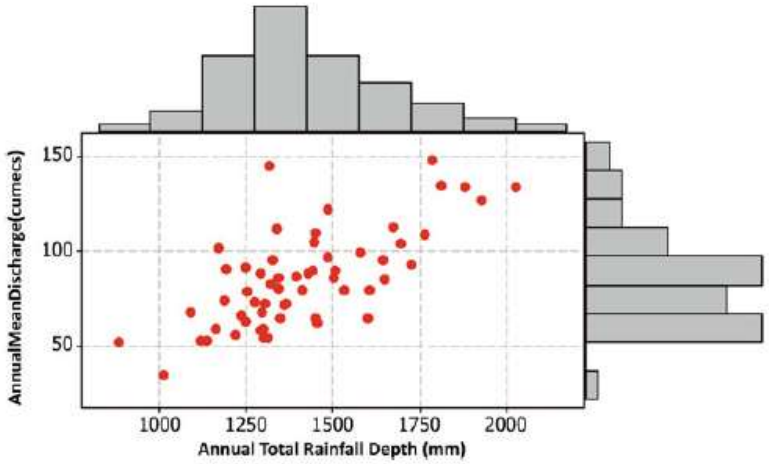
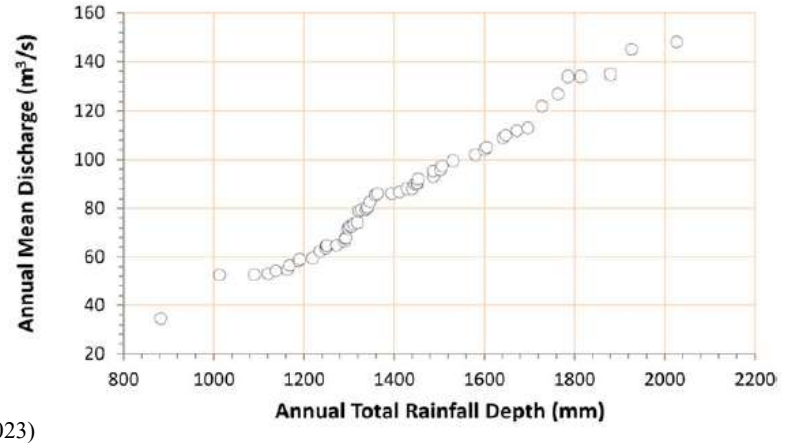


Diagrama Cuantil-Cuantil
Empirical Quantile-Quantile Diagram
Q-Q Plot

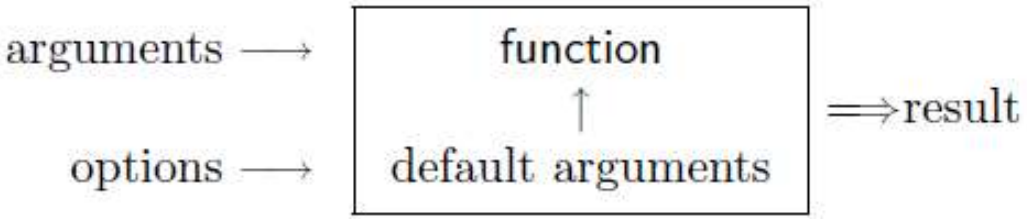


2.4. R-4.3.2 for Windows (32/64 bit) - octubre 2023

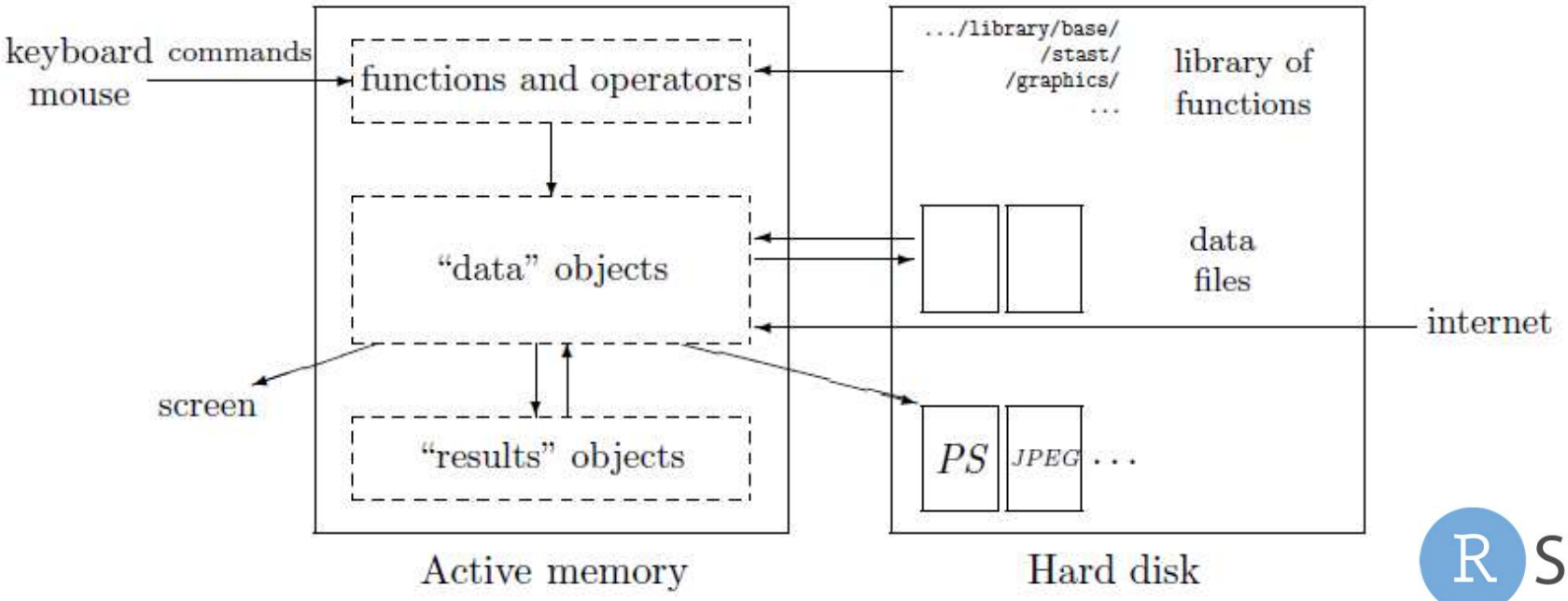
https://cran.r-project.org/bin/windows/base/

- ✓ Sistema para el análisis estadístico y graficación (Ihaka y Gentleman, 1996)
- ✓ Lenguaje interpretado, no es un lenguaje compilado
- ✓ Facil e intuitivo, estrictamente « no es un lenguaje de programación »

posit https://posit.co/download/rstudio-desktop/



Paradis (2002)



RStudio

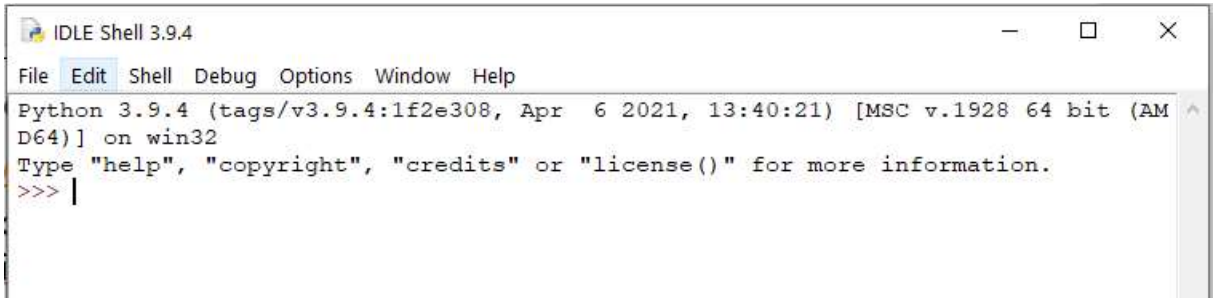
2.5. Python 3.12.0 for Windows (32/64 bit)

https://www.python.org/

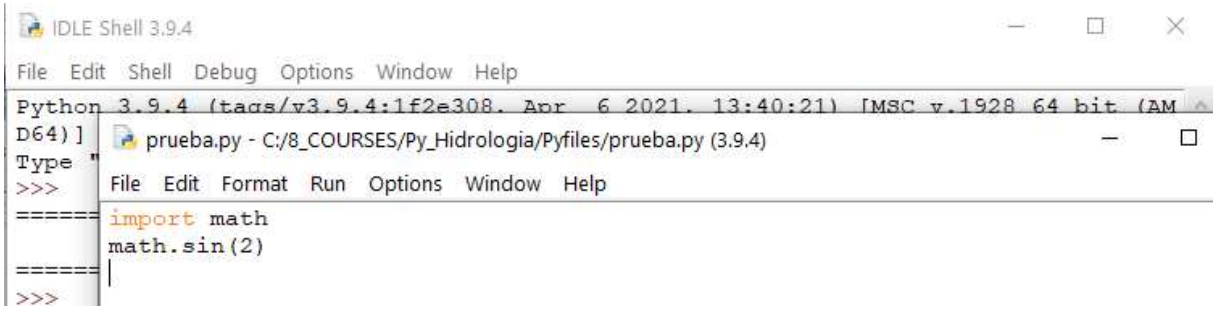
octubre 2023

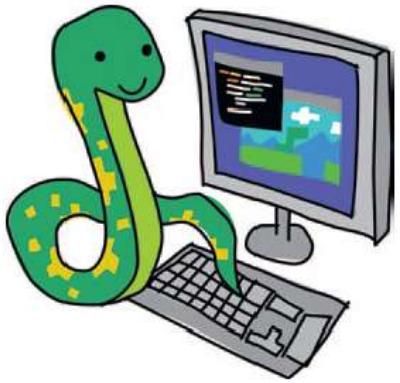
- ✓ Lenguaje de programacion « poderoso » y « rápido » (van Rossum, 1995).
- ✓ Lenguaje interpretado, multiparadigma y casi orientado a objetos.
- ✓ Facil e intuitivo.

*El **IDLE** (Integrated Development and Learning Environment), es el intérprete y permite escribir en Python.*



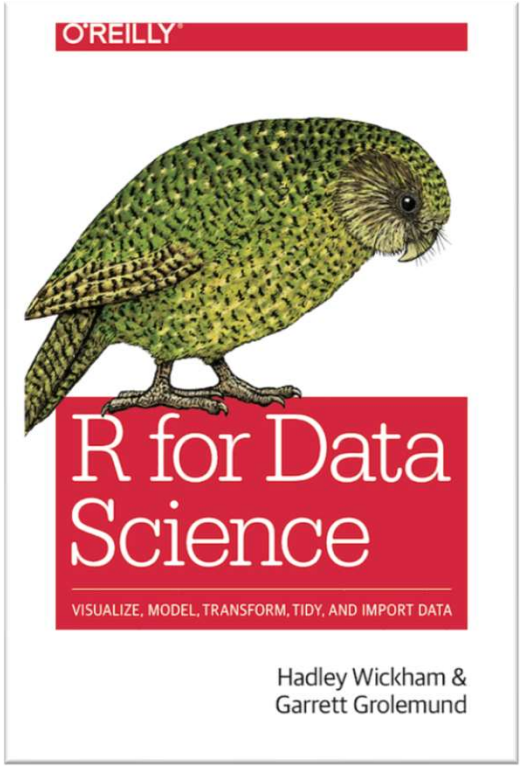
*File – New, para abrir el **editor** de IDLE. También se puede usar el block de notas y guardar el archivo como *.py*





*También desde la **consola de comandos***
***cmd**, usar commandos UNIX "cd" para cambiar la ruta donde se encuentra instalado Python.*

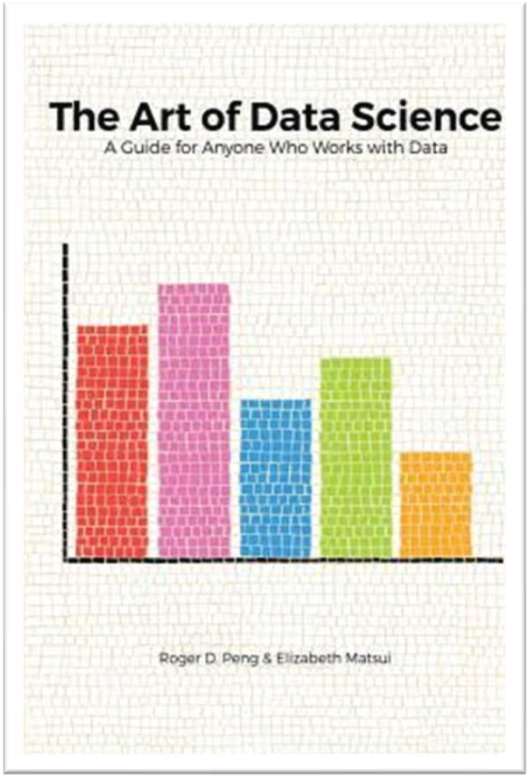
Uso de notebooks



https://r4ds.had.co.nz/

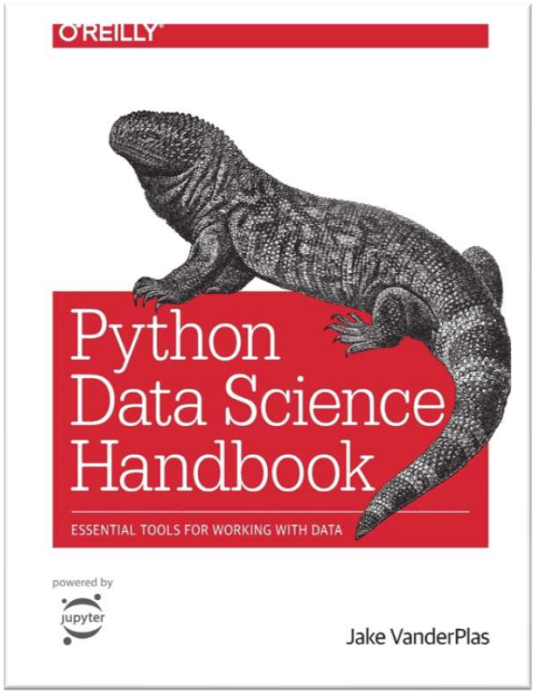
https://es.r4ds.hadley.nz/

https://github.com/jrnold/r4ds-exercise-solutions/blob/master/README.md



https://bookdown.org/rdpeng/artofdatascience/

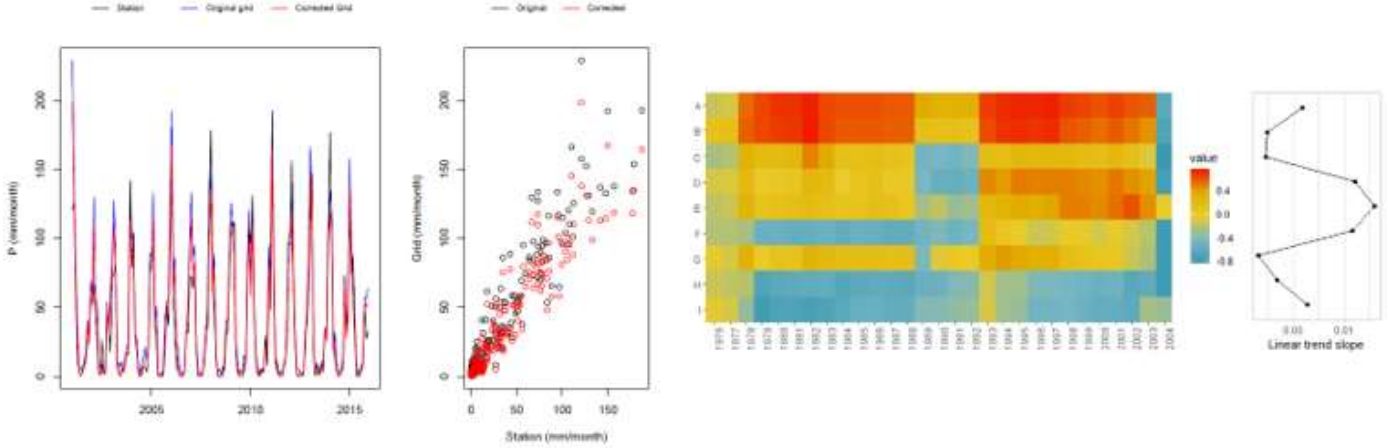
https://github.com/waldronlab/The-Art-of-Data-Science/blob/master/README.md

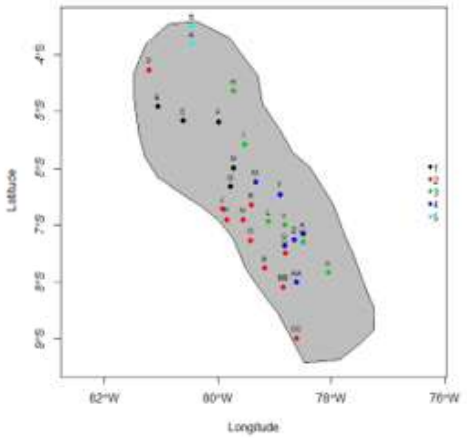


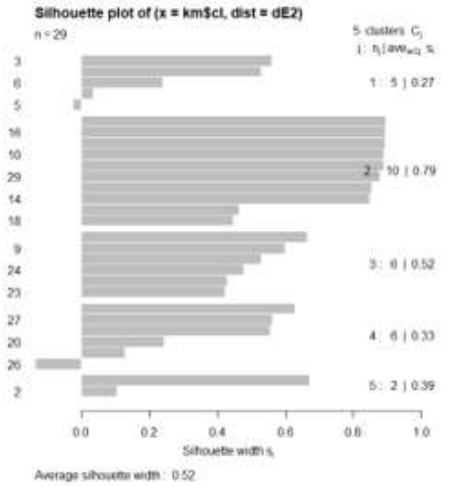
Fuente:
https://jakevdp.github.io/PythonDataScienceHandbook/

Repositorio:
https://github.com/jakevdp/PythonDataScienceHandbook

2.6 Paquetes elaborados en UTEC-CITA

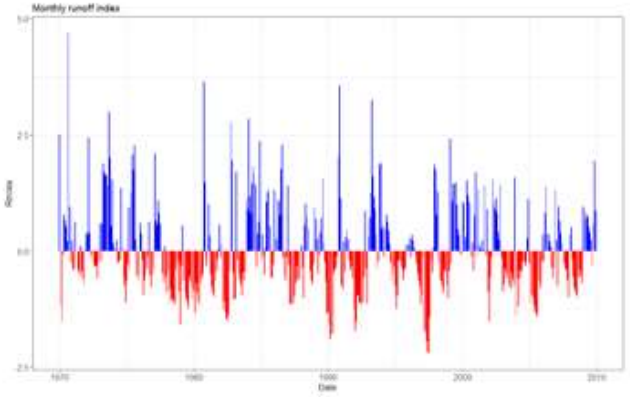


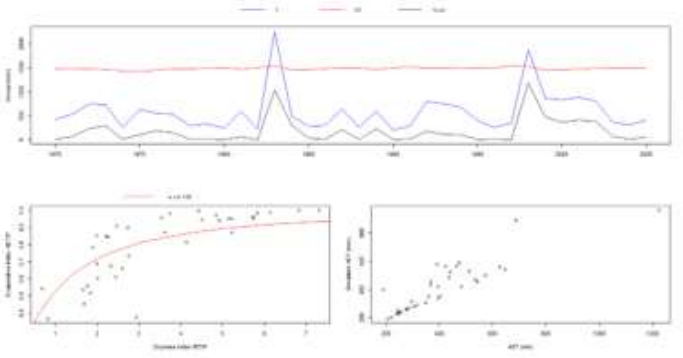


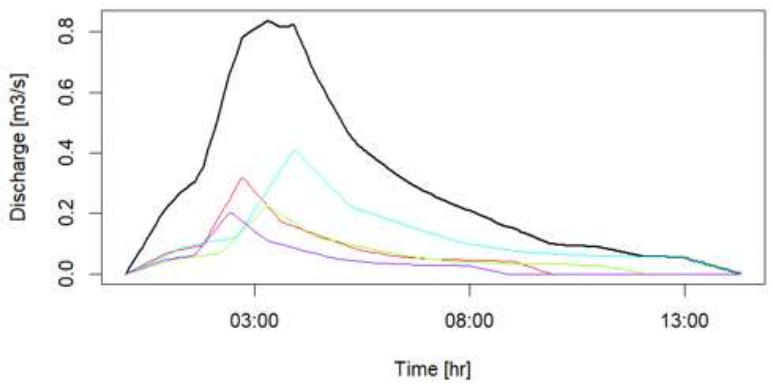


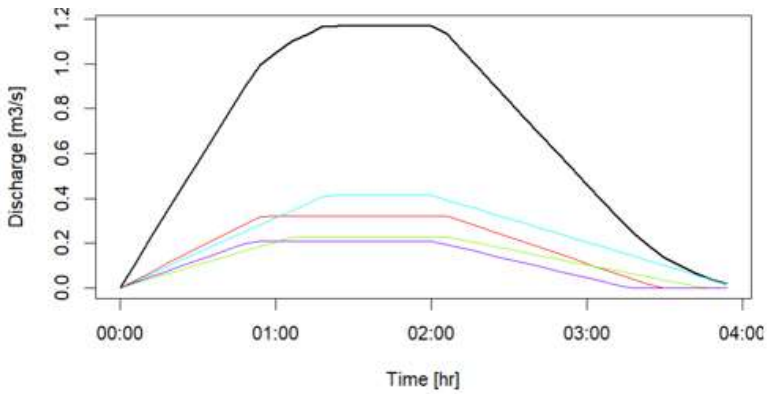
hydRpclim

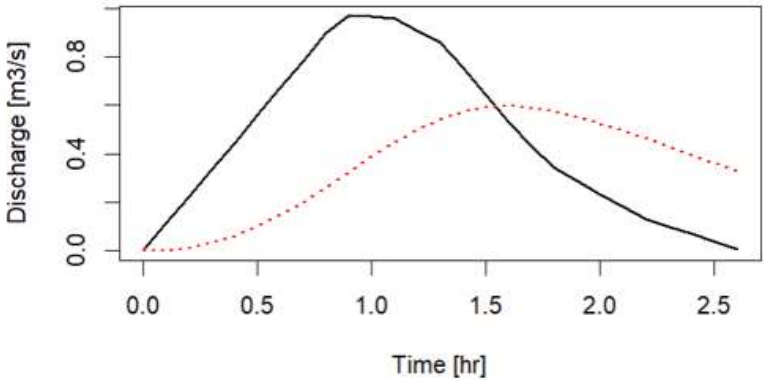
Rau et al (2023)

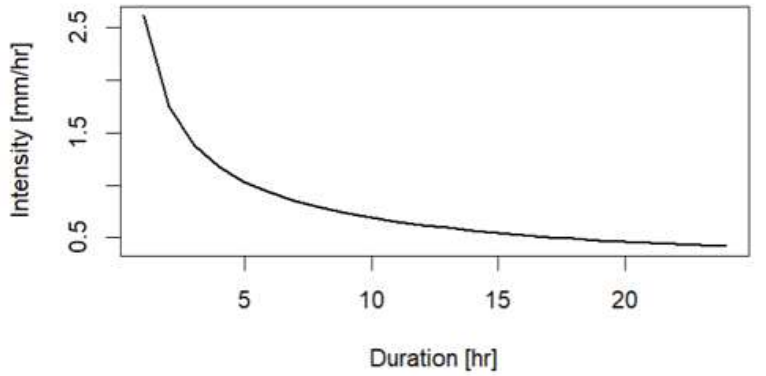


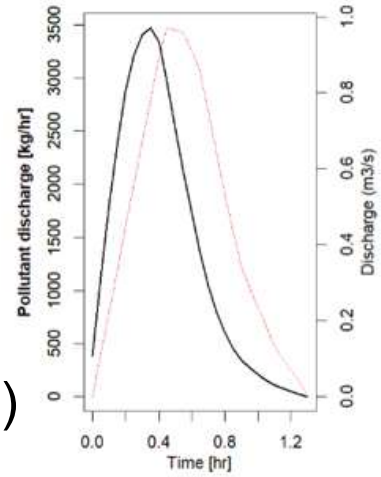


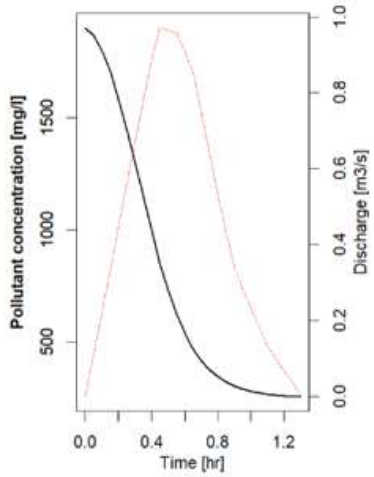


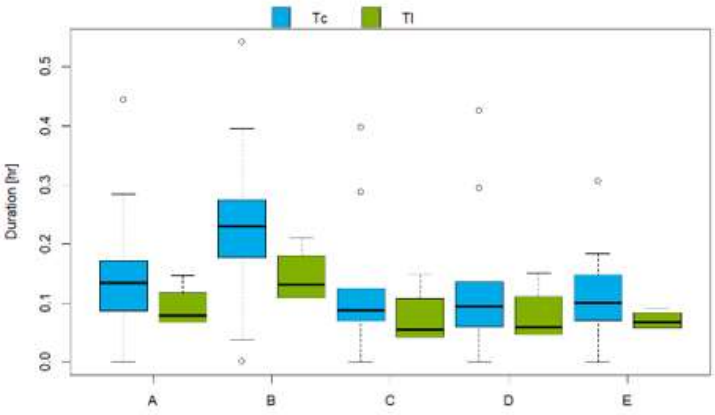












Rau et al (accepted)

A. Procesando una gran cantidad de datos

Ejercicio introductorio

Leer +15 mil datos de Caudales Diarios y convertirlos a Caudales Mensuales. Sabiendo que un caudal mensual equivale al promedio de todos los caudales diarios de dicho mes. Cuidando la ocurrencia de años bisiestos (366 días).

Base de datos CSV
https://github.com/hydrocodes/datascience/blob/main/01-Modulo_intro/diarios.csv

Respuestas A

REVISAR NOTEBOOK COLAB

https://colab.research.google.com/drive/1fiTOXr60elPtoVTnpbpW2atXz5bS8I2Y?usp=sharing

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
data =
pd.read_csv('https://raw.githubusercontent.com/hydrocodes/py.hidrologia/ma
in/diarios.csv', parse_dates = ['fecha'])
print(data.head())
data.fecha = pd.to_datetime(data.fecha)
data.set_index('fecha', inplace=True)
data['Yaroca'].plot()
plt.title("Caudales diarios Yaroca")
plt.ylabel("Q (m3/s)")
# Agregacion
q_m = data.resample('M').agg(pd.Series.mean, skipna=False)
q_m['Yaroca'].plot()
plt.title("Caudales mensuales Yaroca")
plt.ylabel("Q (m3/s)")
```

Referencias

Ihaka R. & Gentleman R. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299-314.

Paradis E. 2002. R for beginners. Institut de Sciences de l'evolution. Université de Montpellier. France

Rau P, Castillon F, Bourrel L. 2023. A tool in R for easy hydroclimatic calculations. Advances in Science, Technology and Innovation (in press).

Rau P, Gutierrez L, Callan N. 2024. A tool in R for handling hydrologic drainage design. (accepted)

Rau P, Bourrel L, Labat D, Melo P, Dewitte B, Frappart F, Lavado W, Felipe O, 2017. Regionalization of rainfall over the Peruvian Pacific slope and coast. International Journal of Climatology 37(1):143-158.

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA

Salas J. 1996. Analysis and modelling of hydrologic time series (in Handbook of Hydrology). McGraw-Hill Education.

Slater L, Thirel G, Harrigan S et al. 2019. Using R in hydrology: a review of recent developments and future directions. Hydrol. Earth Syst. Sci., 23, 2939–2963

van Rossum, G., Drake, F. 2011. The Python Language Reference Manual. Network Theory Ltd.

VanderPlas J. 2017. Python Data Science Handbook. O'Reilly Media.