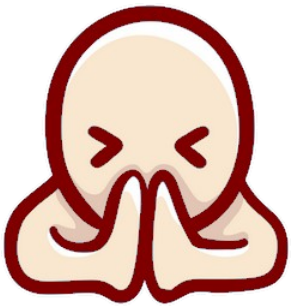
University of the Philippines
LOS BAÑOS

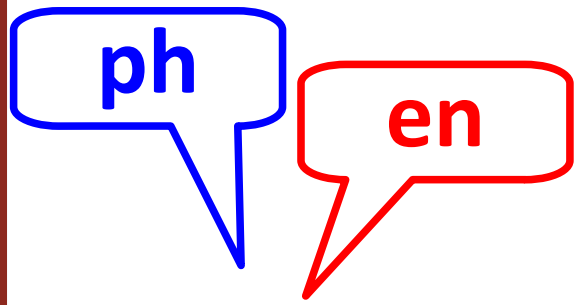
University of the Philippines Data Science Society
University of the Philippines Intelligent Systems Center



University of the Philippines
LOS BAÑOS

From Rainfall Predictions to Crop Simulations:
Model Validation in Climate-Data Pipelines

JADERICK P. PABICO
Research Collaboratory for High Performance Computing
Research Collaboratory for Advanced Intelligent Systems
Institute of Computer Science, UPLB



Workshop for the 5th Philippine Junior Data Science Challenge (PJDSC)

Why simulate crop growth?

- **Claim:** Crop growth simulation models are essential tools for climate resilience
  - **Farmer**'s Perspective
  - **Business** Perspective
  - **Food Security** Perspective
  - **Policy & Research** Perspective

University of the Philippines LOS BAÑOS

Farmer's Perspective

- Climate change = greater volatility
  - What is volatile?
    - Unpredictable planting windows
    - Extreme weather
    - Altered growing season

University of the Philippines LOS BAÑOS

Farmer's Perspective

- From guesswork to Precision
- Crop Model = Virtual Field Laboratory
  - Risk Mitigation
  - Input Optimization (ROI)

University of the Philippines LOS BAÑOS

Business Perspective

- Managing financial and logistical risk
- Businesses that support agriculture operate on forecasts
  - Crop insurance
  - Commodity trading and marketing
  - Warehousing and logistics

University of the Philippines LOS BAÑOS

Food Security Perspective

- Regional stability and early warning
- Early warning system
  - Emergency aid allocation
  - Trade policy adjustments
- Vulnerability mapping

University of the Philippines LOS BAÑOS

Policy & Research Perspective

- Informed and sustainable strategy
- Climate Adaptation Policy
- Optimizing Public Investment
- Setting Goals

University of the Philippines LOS BAÑOS

Why simulate crop growth?

- **Claim:** Crop growth simulation models are essential tools for climate resilience
- The integration of Rainfall Predictions into this framework is what makes the crop model a dynamic tool for resilience rather than just a static calculator of historical averages.

Why simulate crop growth?

- **Claim:** Predictions from crop simulation models are inherently reliable.
  - Mechanistic (process-based)
  - Data-driven
  - Integrative

University of the Philippines LOS BAÑOS

Crop models are reliable

- Mechanistic (process-based)
  - Built on the fundamental laws of physics, chemistry and biology
  - Causality over correlation
  - Mass and energy conservation

University of the Philippines LOS BAÑOS

Crop models are reliable

- Integration and consistency
  - Model integrates multiple complex systems
  - Soil-Plant-Atmosphere Continuum
  - Temporal Reliability (Daily/Hourly Steps)

University of the Philippines LOS BAÑOS

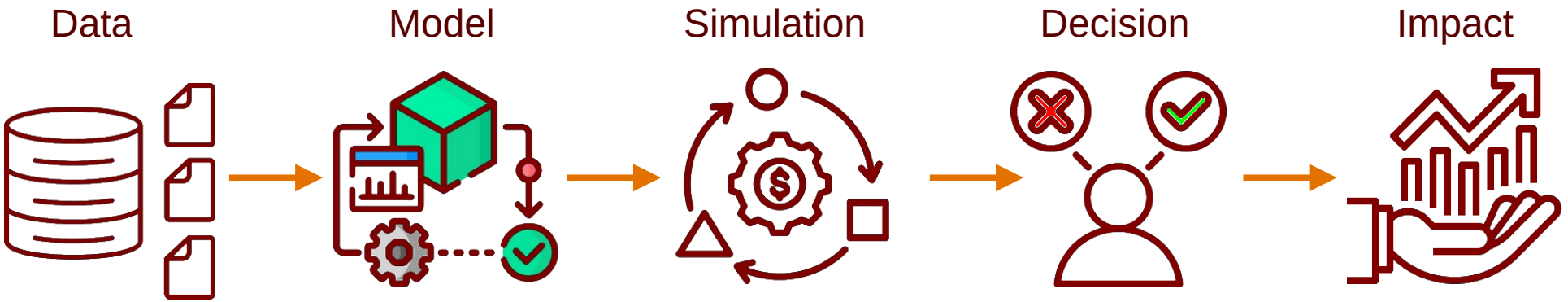
Crop models are reliable

- Scalability and scenario testing
  - Transferability
  - Systematic Uncertainty Quantification

University of the Philippines
LOS BAÑOS

Climate Data Pipeline

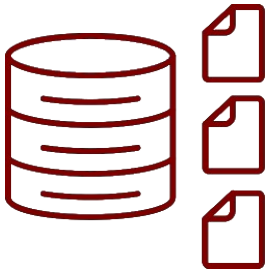
- From data to impact



University of the Philippines LOS BAÑOS

Data: The Foundation of the Pipeline

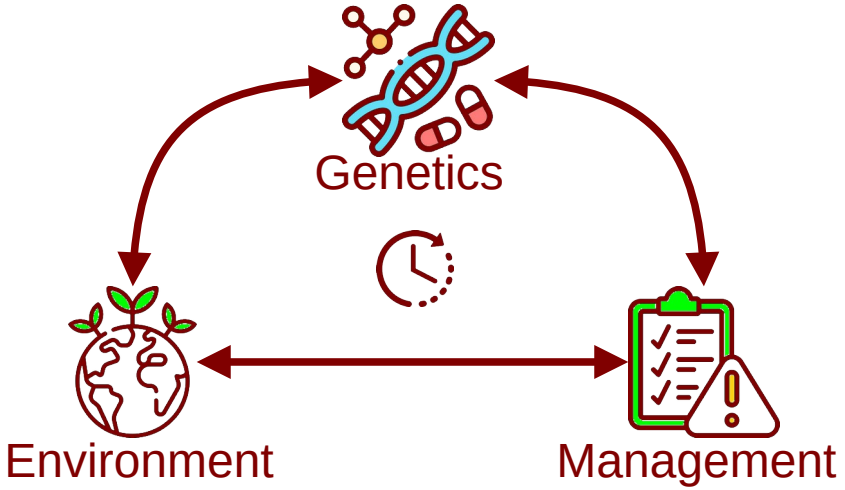
- Collecting, cleaning & processing raw inputs
- Reliability of the pipeline starts here
- Inputs:
  - historical & forecasted weather (temperature, solar radiation, precipitation)
  - Soil characteristics & crop management details
- Challenge: noisy, inconsistent, coarse



University of the Philippines LOS BAÑOS

Model: The Processing Engine

- Mathematical heart of the pipeline
- Process-based computational tool
- Quantifies the interaction of:



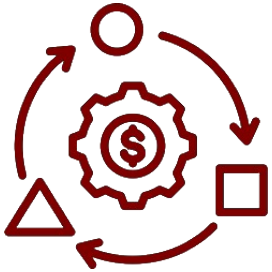
University of the Philippines LOS BAÑOS

Model: The Processing Engine

- Function:
  - uses differential equations
  - Simulate daily processes (photosynthesis, transpiration, biomass accumulation, phenological development)
- Validation rule:
  - Calibration
  - Validation

University of the Philippines LOS BAÑOS

Simulation: Inputs to Outcomes



- Running the validated model under various scenarios
- Scenario Testing: Thousands of runs
- Output: Probabilistic distribution

University of the Philippines LOS BAÑOS

Decision: Translating Risk into Action

- Results are translated into a concrete, actionable choice for the stakeholder.
- **Farmer**: When to plant / How much fertilizer
- **Business/Policy**:
  - setting crop insurance premiums
  - forecasting commodity prices
  - directing government resources for drought relief.

University of the Philippines LOS BAÑOS

Impact: The Real-World Outcome

- Economic
- Environmental
- Social/Resilience

The pipeline demonstrates that validation is not just a scientific step; it is an **ethical requirement** to ensure that the model produces reliable Decisions that lead to positive Impact.

Why this Matters?

- SCIENCE ADVANCES | RESEARCH ARTICLE

ATMOSPHERIC SCIENCE

Global concurrent climate extremes exacerbated by anthropogenic climate change

Sha Zhou[1,2]*, Bofu Yu[3], Yao Zhang[4]

Increases in concurrent climate extremes in different parts of the world threaten the ecosystem and our society. However, spatial patterns of these extremes and their past and future changes remain unclear. Here, we develop a statistical framework to test for spatial dependence and show widespread dependence of temperature and precipitation extremes in observations and model simulations, with more frequent than expected concurrence of extremes around the world. Historical anthropogenic forcing has strengthened the concurrence of temperature extremes over 56% of 946 global paired regions, particularly in the tropics, but has not yet significantly affected concurrent precipitation extremes during 1901–2020. The future high-emissions pathway of SSP585 will substantially amplify the concurrence strength, intensity, and spatial extent for both temperature and precipitation extremes, especially over tropical and boreal regions, while the mitigation pathway of SSP126 can ameliorate the increase in concurrent climate extremes for these high-risk regions. Our findings will inform adaptation strategies to alleviate the impact of future climate extremes.

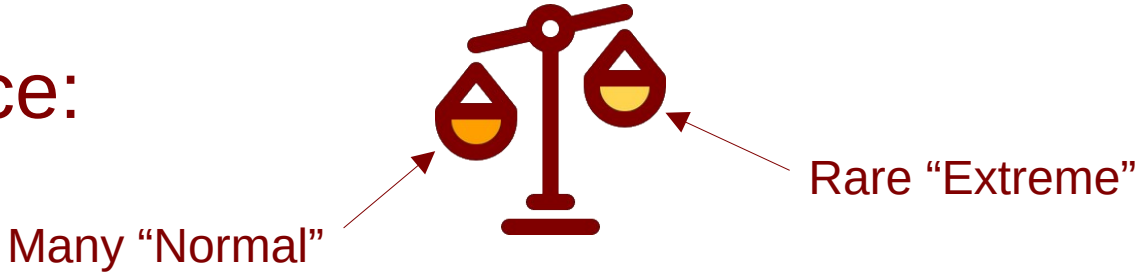
Sha Zhou S, B Yu & Y Zhang. 2023. *Global concurrent climate extremes exacerbated by anthropogenic climate change*. **Science Advances** 9(10):eabo1638.

Typhoon → Rainfall Forecasting → Crop Management Decision

University of the Philippines LOS BAÑOS

Climate Data Realities: Imbalance & Noise

- Imbalance:



- Noisy:
  - sensor errors
  - missing values
  - measurement uncertainty

Real-World Example: Imbalance & Noise

- Taal Lake Fishkill

SOCIAL SENSOR AS AN EMERGING REMOTE SENSING TOOL AND ITS APPLICATION TO SENSING TAAL LAKE FISHKILL

Jaderick P. Pabico,[1] Arnold R. Salvacion[2] and Damasa B. Magcale-Macandog[3]

[1]Research Collaboratory for Advanced Intelligent Systems, Institute of Computer Science, University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: jppabico@up.edu.ph

[2]Community and Environmental Resource Planning, College of Human Ecology University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: arsalvacion@up.edu.ph

[3]Ecoinformatics Laboratory, Institute of Biological Sciences, College of Arts and Sciences University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: dbmmacandog@up.edu.ph

KEYWORDS: Social sensor, emerging remote sensing tool, Twitter, Taal Lake fishkill

- Imbalance:

Many “No Fishkill”

Rare “Fishkill”

Real-World Example: Imbalance & Noise

- Taal Lake Fishkill

SOCIAL SENSOR AS AN EMERGING REMOTE SENSING TOOL AND ITS APPLICATION TO SENSING TAAL LAKE FISHKILL

Jaderick P. Pabico,[1] Arnold R. Salvacion[2] and Damasa B. Magcale-Macandog[3]

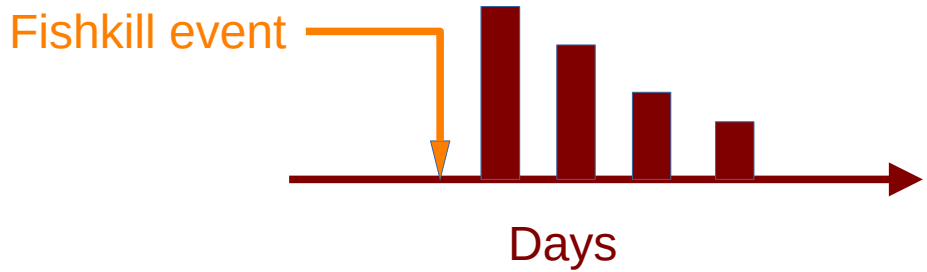
[1]Research Collaboratory for Advanced Intelligent Systems, Institute of Computer Science, University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: jppabico@up.edu.ph

[2]Community and Environmental Resource Planning, College of Human Ecology University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: arsalvacion@up.edu.ph

[3]Ecoinformatics Laboratory, Institute of Biological Sciences, College of Arts and Sciences University of the Philippines Los Baños, College 4031, Laguna, Philippines
Email: dbmmacandog@up.edu.ph

KEYWORDS: Social sensor, emerging remote sensing tool, Twitter, Taal Lake fishkill

- Missing Data:



Pabico JP, AR Salvacion & DB Magcale-Macandog. 2015. *Social sensor as an emerging remote sensing tool and its application to sensing Taal Lake fishkill.* **The 36th Asian Conference on Remote Sensing (ACRS 2015)**, Crowne Plaza Manila Galleria, Quezon City, 19-23 October 2015.

Real-World Example: Imbalance & Noise

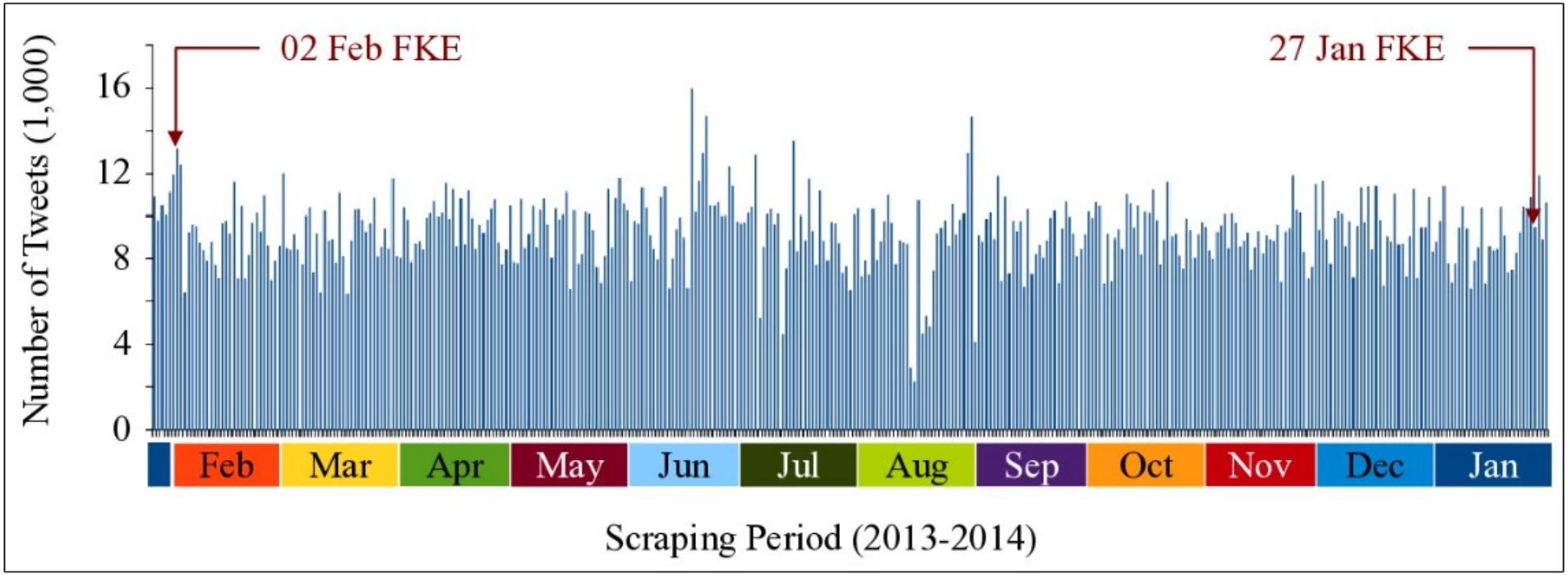


Figure 1. The daily total number of tweets collected by pScraper.

Pabico et al. 2015

University of the Philippines
LOS BAÑOS

Real-World Example: Imbalance & Noise

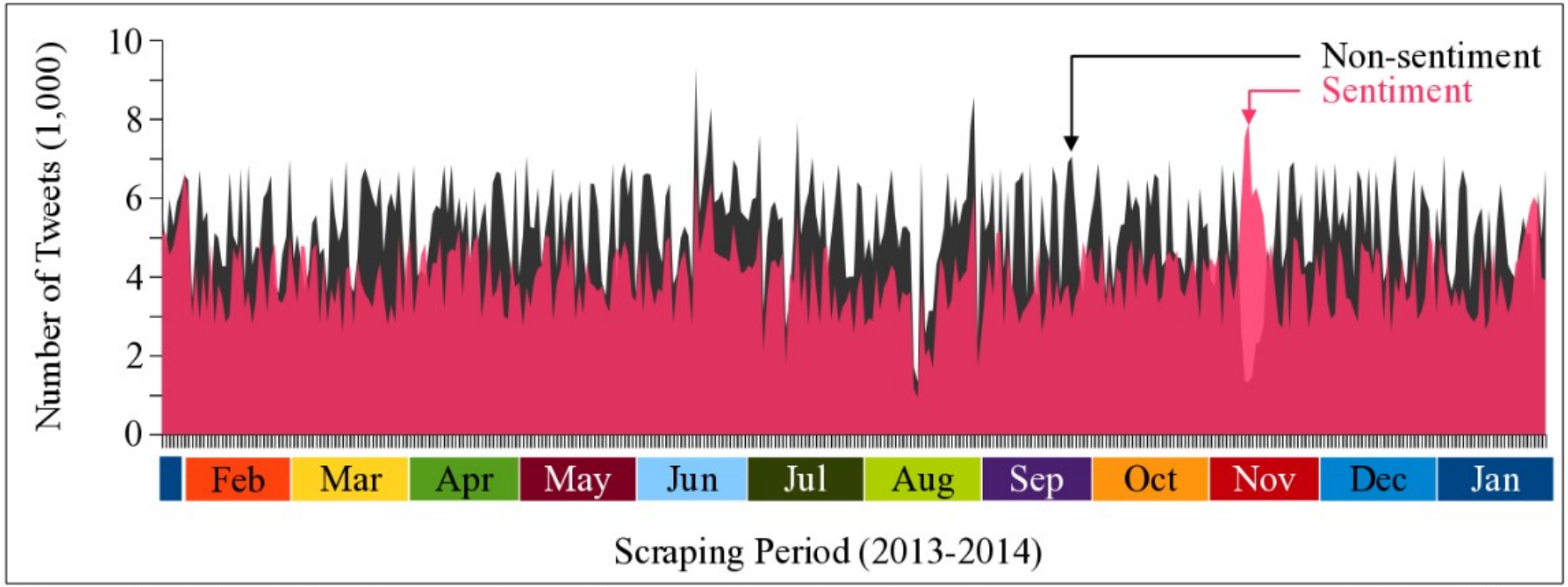


Figure 2. The daily total number of non-sentiment and sentiment tweets as computed by pClass. Black bars are non-sentiment and red bars are sentiment tweets.

Pabico et al. 2015

Real-World Example: Imbalance & Noise

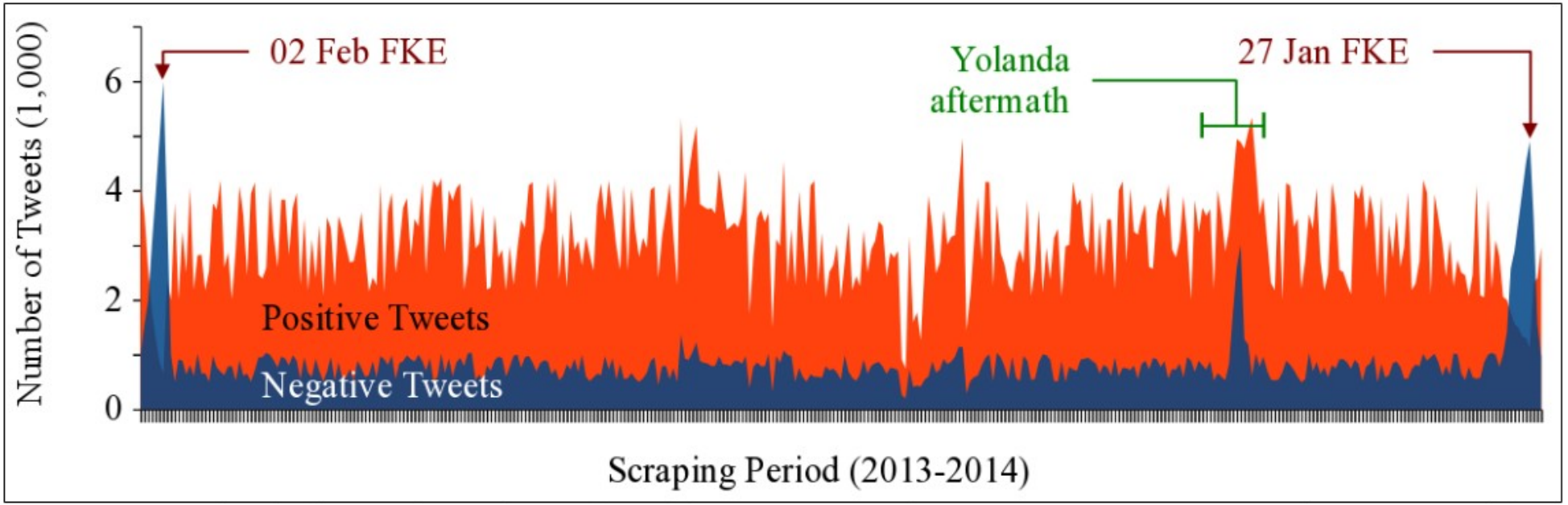


Figure 3. The daily total number of positive and negative CSP tweets. Orange bars are positive CSP tweets and blue bars are negative CSP tweets.

*CSP = Contextual Sentiment Polarity

Pabico et al. 2015

Real-World Example: Imbalance & Noise

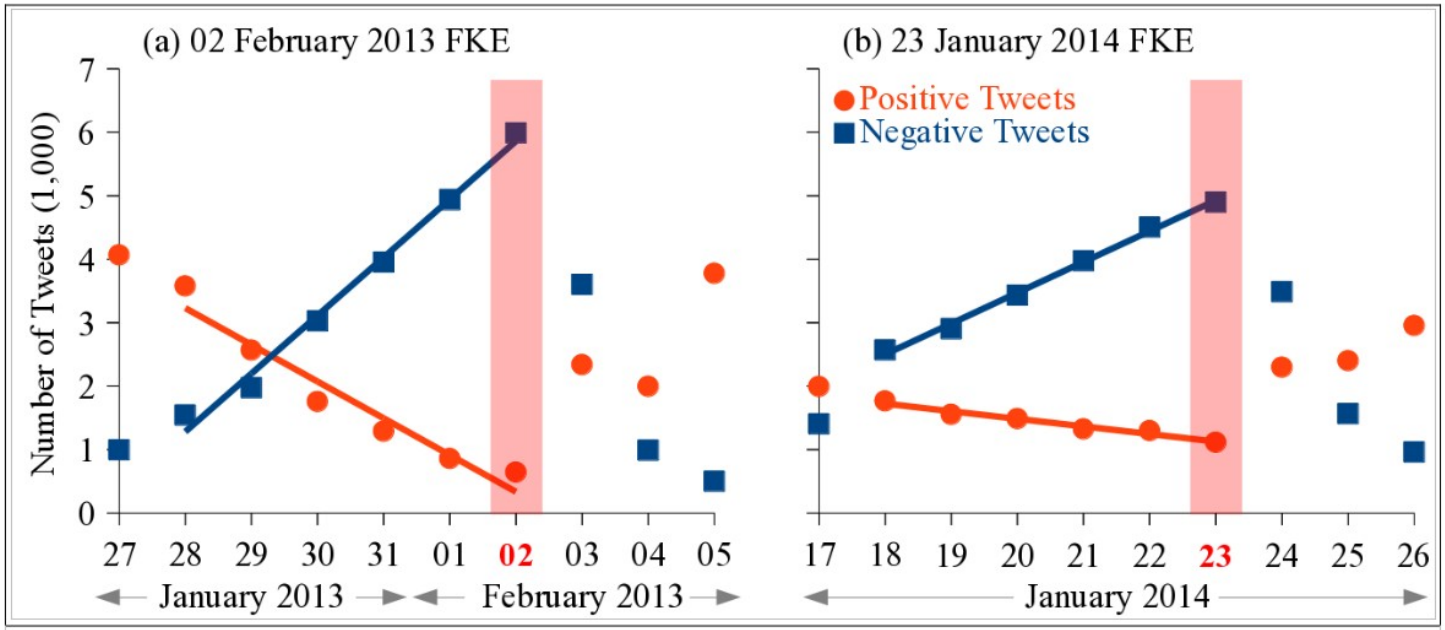


Figure 4. The respective linear trends of the negative and positive sentiment tweets around (a) the 02 February FKE, and (b) the 16 January 2014 FKE.

Pabico et al. 2015

University of the Philippines LOS BAÑOS

Imbalance Data: Why it matters for modeling

- Standard accuracy mis-leads under imbalance classes
- Need alternate metrics (precision/recall, F1, PR-curve)

remote sensing

MDPI

Article

A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications

Tajul Miftahushudur [1,2], Halil Mertkan Sahin [2], Bruce Grieve [2] and Hujun Yin [2,*]

[1] Research Centre for Telecommunication, National Research and Innovation Agency (BRIN), Bandung 40135, Indonesia; tajul.miftahushudur@postgrad.manchester.ac.uk or mtaj001@brin.go.id
[2] Department of Electrical and Electronic Engineering, The University of Manchester, Manchester M13 9PL, UK; halil.sahin@manchester.ac.uk (H.M.S.); bruce.grieve@manchester.ac.uk (B.G.)
* Correspondence: hujun.yin@manchester.ac.uk

Miftahushudur T, HM Sahin, B Grieve & H Yin. 2025. *A Survey of Methods for Addressing Imbalance Data Problems in Agriculture Applications*. **Remote Sensing** 17(3):454.

Imbalance Data: Why it matters for modeling

- Resampling Methods:

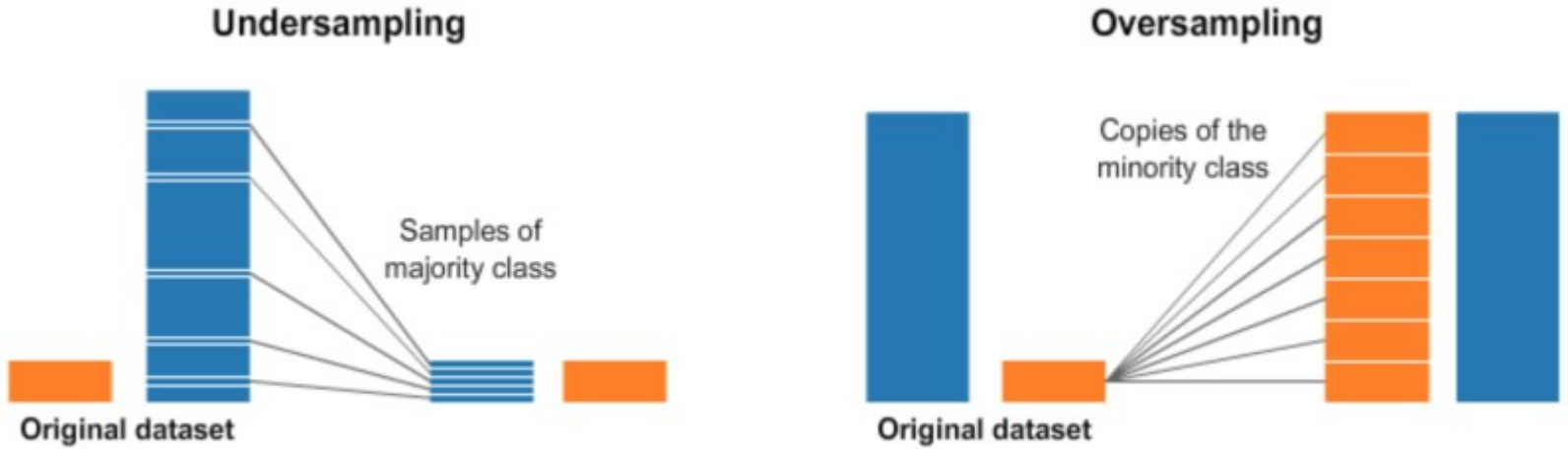


Figure 1. Difference between undersampling and oversampling [102]. (Blue and orange colors represent majority and minority classes, respectively).

Miftahushudur et al., 2025.

Imbalance Data: Why it matters for modeling

- Undersampling Methods:
  - Random undersampling (RUS)
  - Edited nearest neighbor
  - Neighborhood cleaning rule
  - Tomek-Links
  - Cluster-based Oversampling

Miftahushudur et al., 2025.

University of the Philippines LOS BAÑOS

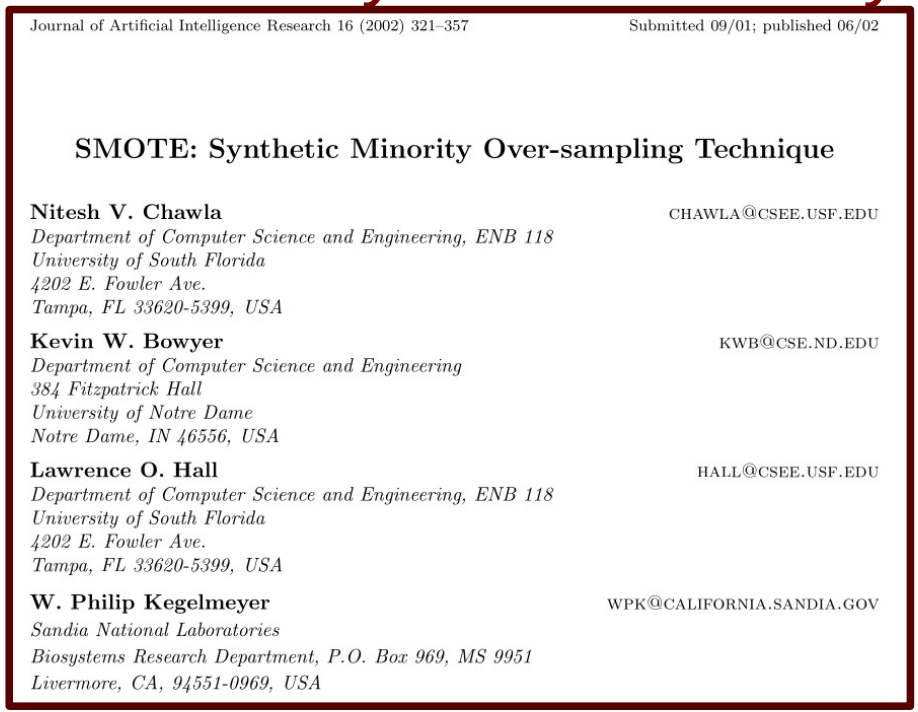
Imbalance Data: Why it matters for modeling

- Undersampling Methods:
  - Random oversampling (ROS)
  - SMOTE: synthetic minority oversampling technique
  - ADASYN
  - Borderline SMOTE
  - SL-SMOTE
  - K-Means SMOTE
  - SVM SMOTE

Miftahushudur et al., 2025.

SMOTE

- SMOTE: synthetic minority oversampling technique



Journal of Artificial Intelligence Research 16 (2002) 321–357 Submitted 09/01; published 06/02

SMOTE: Synthetic Minority Over-sampling Technique

Nitesh V. Chawla CHAWLA@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

Kevin W. Bowyer KWB@CSE.ND.EDU
Department of Computer Science and Engineering
384 Fitzpatrick Hall
University of Notre Dame
Notre Dame, IN 46556, USA

Lawrence O. Hall HALL@CSEE.USF.EDU
Department of Computer Science and Engineering, ENB 118
University of South Florida
4202 E. Fowler Ave.
Tampa, FL 33620-5399, USA

W. Philip Kegelmeyer WPK@CALIFORNIA.SANDIA.GOV
Sandia National Laboratories
Biosystems Research Department, P.O. Box 969, MS 9951
Livermore, CA, 94551-0969, USA

Chawla NV, KW Bowyer, LO Hall & W Philip Kegelmeyer. 2002. *SMOTE: Synthetic Minority Over-sampling Technique*. **Journal of Artificial Intelligence Research** 16(2002):321-357.

SMOTE in 5 easy steps

- **(1/5) Identify the Minority Class**
  - Determine which class has fewer samples (e.g., “extreme rainfall days” in your dataset).
  - Let’s call it Class 1 (minority); the other is Class 0 (majority).

Chawla et al., 2002

University of the Philippines LOS BAÑOS

SMOTE in 5 easy steps

- **(2/5) Select a Minority Sample and Its Nearest Neighbors**
  - For each sample x_i in the minority, find its **k-nearest neighbors** among other minority samples (usually $k = 5$)
  - Distance is computed in feature space (e.g., euclidean)

Chawla et al., 2002

SMOTE in 5 easy steps

- **(3/5) Randomly Choose a Neighbor**
  - Randomly select one of the *k* neighbors, x_{nn}.
  - The pair (x_i, x_{nn}) will be used to synthesize a new sample.

Chawla et al., 2002

University of the Philippines LOS BAÑOS

SMOTE in 5 easy steps

- **(4/5) Create a Synthetic Sample**
  - Generate a new, synthetic sample along the line segment between x_i and x_{nn}.

    $x_{new} = x_i + \delta(x_i - x_{nn})$, δ = random(0,1)

  - This produces a *convex combination* of existing minority samples — not a copy, but an *interpolation*.

Chawla et al., 2002

University of the Philippines LOS BAÑOS

SMOTE in 5 easy steps

- **(5/5) Repeat Until the Minority Class Is Balanced**
  - Repeat the process for randomly selected minority samples until the minority class size equals (or reaches a chosen ratio of) the majority class.
  - The resulting dataset has **synthetic but plausible** minority examples distributed in feature space, reducing class imbalance.

Chawla et al., 2002

Noise & Data Quality Issues

- Errors in rainfall measurement propagate downstream
- Examples of noise: gauge errors, satellite retrievals, interpolation bias

ELSEVIER

Ecological Modelling 157 (2002) 1–21

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

The impact of potential errors in rainfall observation on the simulation of crop growth, development and yield

Alexandre B. Heinemann [1], Gerrit Hoogenboom *, Bogdan Chojnicki [2]

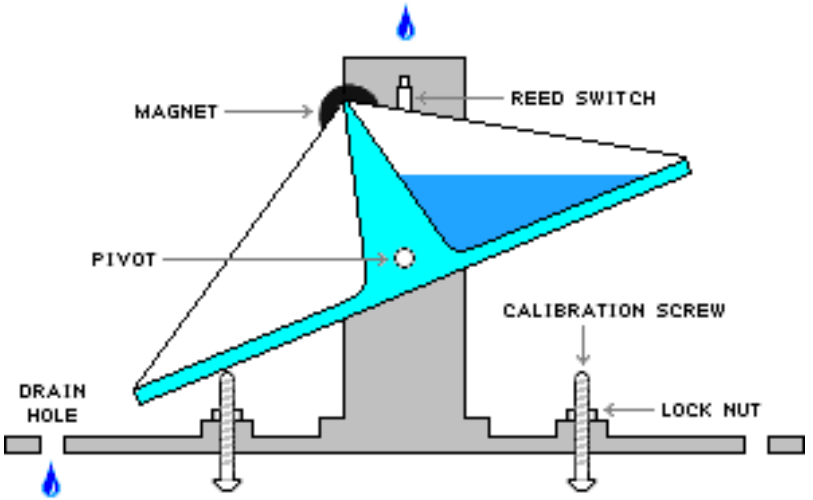
Department of Biological and Agricultural Engineering, The University of Georgia, 1109 Experiment Street, Griffin, GA 30223-1797, USA

Received 16 October 2001; received in revised form 5 April 2002; accepted 16 April 2002

Heinemann AB, G Hoogenboom & B Chojnicki. 2002. *The impact of potential errors in rainfall observation on the simulation of crop growth, development and yield*. **Ecological Modelling** 157(1):1-21.

Noise & Data Quality Issues

- Most Automated Weather Stations (AWS) use tipping-bucket rain gauge (TBRG)
- Errors of TBRG:
  - Aerodynamic effects
  - Wetting losses
  - Design and operation of the tipping-bucket sensor



Heinemann et al. 2002

University of the Philippines LOS BAÑOS

What did the authors do:

- Evaluated how inaccuracies in measured rainfall (both random and systematic) affect the outputs of crop simulation models.
- Centered on error propagation — how deviations in rainfall input changed simulated variables like soil moisture, evapotranspiration, phenology, and yield.
- Explore how temporal distribution of rainfall affects outcomes — i.e., missing or misplacing rainfall events early or late in the growing season causes nonlinear effects on growth and yield.
- Temporal imbalance (in rainfall occurrence over time) is a hydrological heterogeneity issue, not a data-class imbalance problem.

Heinemann et al. 2002

University of the Philippines LOS BAÑOS

Key findings:

- Rainfall observation errors significantly influence crop simulation outputs.
- Model sensitivity depends on the type of crop and water-limitation context.
- Temporal distribution of rainfall errors matters more than total rainfall errors.

Heinemann et al. 2002

University of the Philippines LOS BAÑOS

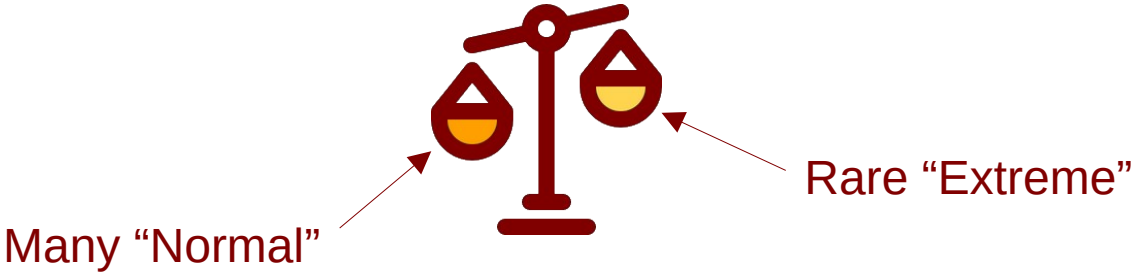
Key findings:

- Spatial representativeness of rainfall data is crucial.
- Systematic bias in rainfall data compounds over the growing season.
- Crop model calibration cannot fully offset poor rainfall data.
- Practical implication

Heinemann et al. 2002

University of the Philippines LOS BAÑOS

What “Imbalanced Data” means in DS & ML



- Effect: causes the model to overfit the “normal” and underdetect the rare yet critical “extremes”
- In rainfall datasets, extreme events (storms, floods) are rare, while normal or dry days dominate.

What Heinemann et al. (2002) studied

- Examined what happens when rainfall inputs contain **systematic bias** (e.g., consistent underestimation) or **random noise** (e.g., missing or misplaced rain events).
- These errors caused:
  - Miscalculated soil-water balance
  - Shifted water-stress timing
  - Reduced or inflated simulated yield

“Event Imbalance” as the Hidden Bias

- Heinemann et al. (2002) results implicitly reveal that:
  - **Rainfall time series are inherently imbalanced**: only a small fraction of days contribute most of the total seasonal rainfall.
  - **Extreme rainfall events — though rare — have disproportionate influence** on crop growth, especially in rainfed systems.
  - When measurement errors occur on these rare but critical days, the **downstream model impact** is far greater than if errors occurred on ordinary days.
- In DS terms: These “rare rainfall events” are the minority class, and the model’s response (yield simulation) is highly sensitive to misclassification of that class.

University of the Philippines LOS BAÑOS

Why It Matters for Model Validation

- This mapping teaches that:
  - Treating rainfall data errors uniformly (as random noise) ignores their imbalanced importance — some errors matter far more than others.
  - In data-driven climate AI models, **imbalanced sampling of extremes** can reproduce exactly the same issue: underprediction of floods, droughts, or water stress periods.
- Hence, model validation must be weighted toward the accurate detection and representation of **rare but high-impact events**.

University of the Philippines LOS BAÑOS

Practical Stuff:

Type of Data Bias	Equivalent in Simulation Context	Consequence on Crop Model
Class imbalance (DS)	Rainfall dominated by normal days	Model underrepresents extremes
Minority class misclassification	Misplaced or missing heavy rainfall events	Underestimated yield or false water stress
Oversampling / SMOTE (DS fix)	Rainfall downscaling or stochastic weather generator	Adds synthetic extremes to correct imbalance
Validation metric shift (accuracy → recall)	Crop-model sensitivity analysis focused on dry spells	Emphasizes performance under stress

University of the Philippines LOS BAÑOS

Reflections

- In climate–agriculture modeling, imbalance doesn't always show up as class counts — sometimes it hides as **temporal or spatial imbalance** in the events that matter most.
- Whether we're predicting rainfall or simulating yields, our responsibility is the same: **ensure that rare, high-impact conditions are represented, validated, and trusted**.
- When rare events are underrepresented or misrecorded, simulation models — just like DS models — learn/output an **incomplete story of reality**.

What was found by a similar study?

- 

Article

Effects of Different Spatial Precipitation Input Data on Crop Model Outputs under a Central European Climate

Sabina Thaler [1,2,*], Luca Brocca [3], Luca Ciabatta [3], Josef Eitzinger [1], Sebastian Hahn [4] and Wolfgang Wagner [4]

1 Institute of Meteorology, University of Natural Resources and Life Sciences (BOKU), Gregor-Mendel-Straße 33, 1180 Vienna, Austria; josef.eitzinger@boku.ac.at
2 CzechGlobe—Global Change Research Institute CAS, Belidla 986, 4a, 603 00 Brno, Czech Republic
3 Research Institute for Geo-Hydrological Protection, National Research Council, Via della Madonna Alta 126, 06128 Perugia, Italy; luca.brocca@irpi.cnr.it (L.B.); luca.ciabatta@irpi.cnr.it (L.C.)
4 Department of Geodesy and Geoinformation, Vienna University of Technology (TU Wien), Gußhausstraße 27–29, 1040 Vienna, Austria; sebastian.hahn@geo.tuwien.ac.at (S.H.); wolfgang.wagner@geo.tuwien.ac.at (W.W.)
* Correspondence: sabina.thaler@boku.ac.at; Tel.: +43-1-47654-81420

Received: 30 May 2018; Accepted: 23 July 2018; Published: 26 July 2018

Thaler S, L Brocca, L Ciabatta, J Eitzinger, S Hahn & W Wagner. 2018. *Effects of Different Spatial Precipitation Input Data on Crop Model Outputs under a Central European Climate*. **Atmosphere** 9(8):290.

University of the Philippines LOS BAÑOS

Key findings:

- Spatial variability of precipitation data significantly affects crop simulation results.
- High-resolution, locally measured precipitation yields the most reliable crop-model outputs.
- Interpolation smooths rainfall extremes, leading to underestimation of water stress events.

Thaler et al. 2018

University of the Philippines LOS BAÑOS

Key findings:

- Satellite-based rainfall products can misrepresent local rainfall distribution.
- Model sensitivity depends on climate regime and crop type.
- Temporal resolution is as important as spatial accuracy.
- Quantitative impact
- Practical implication

Thaler et al. 2018

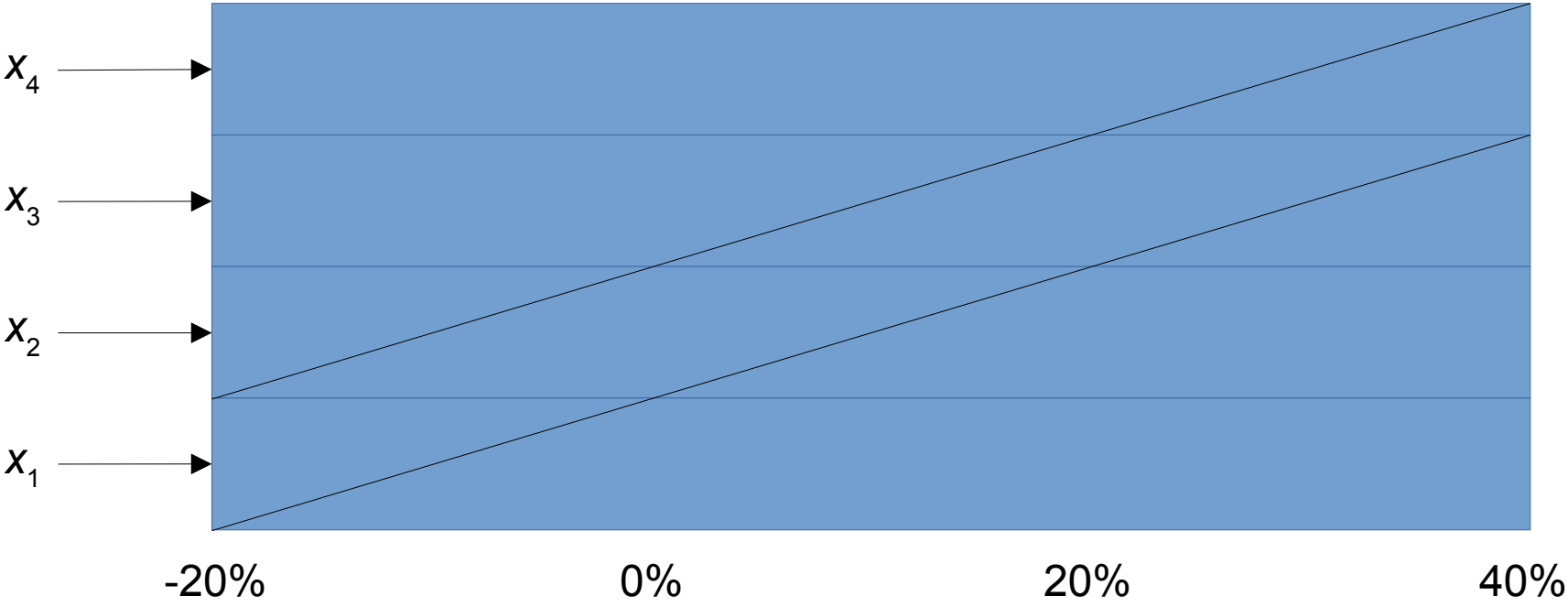
University of the Philippines LOS BAÑOS

Bonus: Encoding Quantitative Variables

- Continuous valued: Any number within a range of values
- Periodic values: Values repeat periodically
- **Interpolation Representation**
- Example:
  - Company Growth Rate (ranges from -20% to 40%)

Pabico JP. 1996. **Fundamentals of Neural computing**. Institute of Computer Science, UPLB

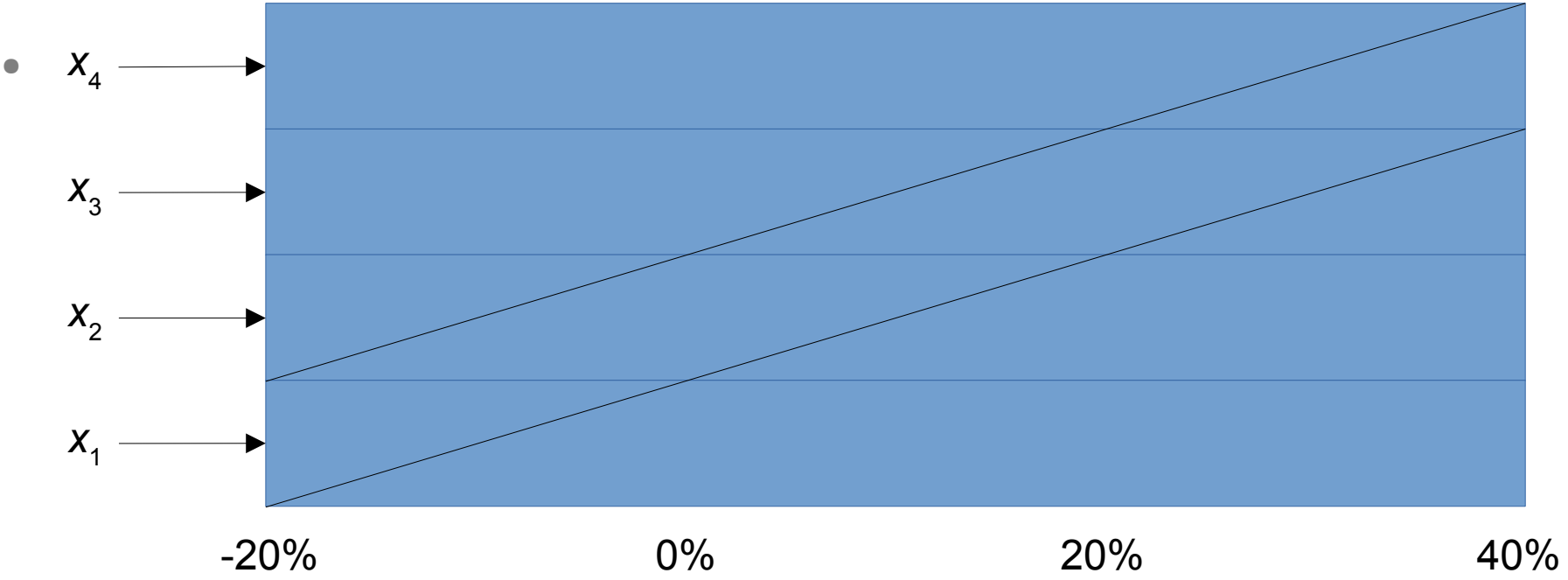
Interpolation Representation



Values of x_1, x_2, x_3, and x_4 are continuous between 0 and 1.

Pabico, 1996.

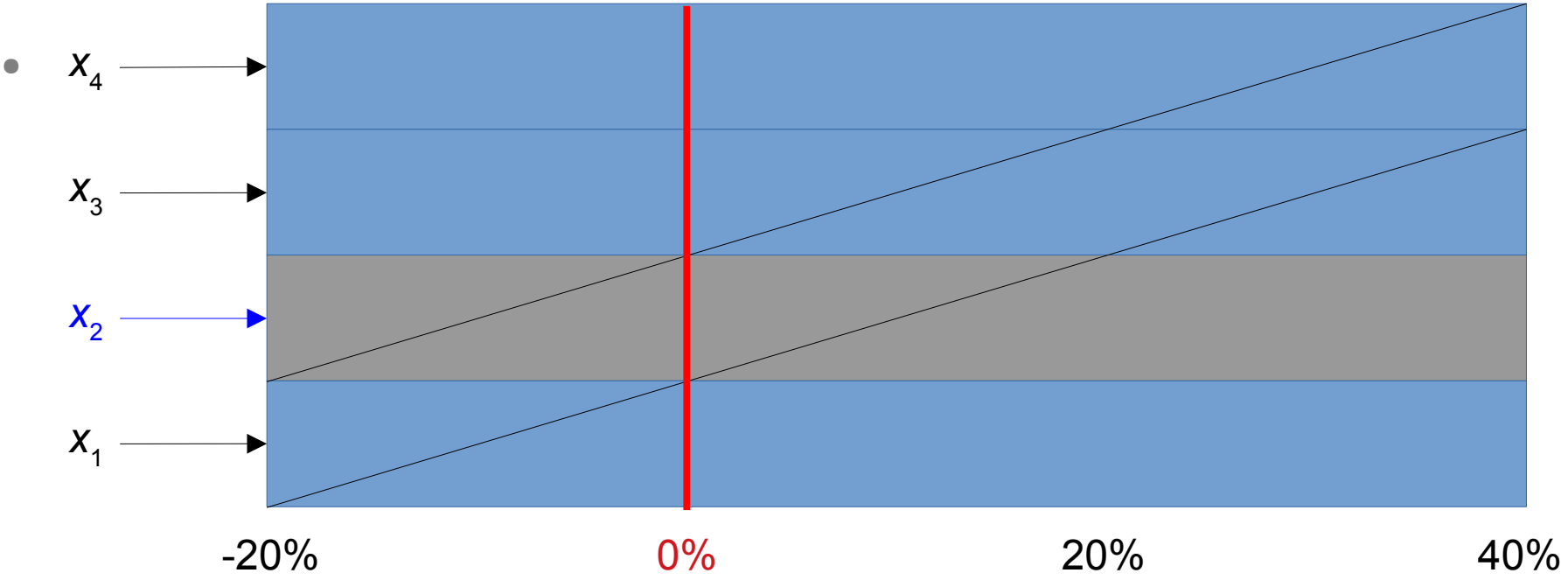
Interpolation Representation



Variable x_1 represents -20%, x_2 0%, x_3 20% and x_4 40% growth

Pabico, 1996.

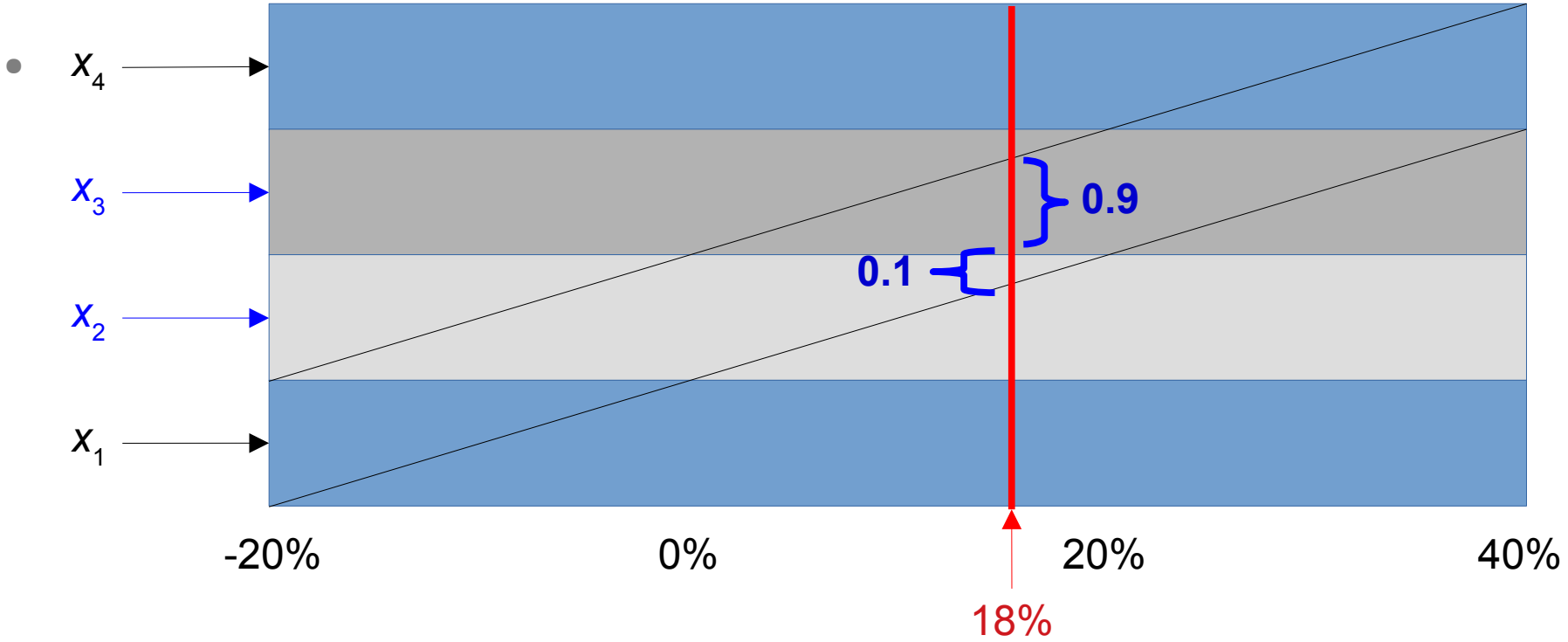
Interpolation Representation



If growth is exactly 0%, x_2 value = 1

Pabico, 1996.

Interpolation Representation



If growth is between any of the numbers, example, 18% growth means x_2 has 0.1 and x_3 has 0.9.

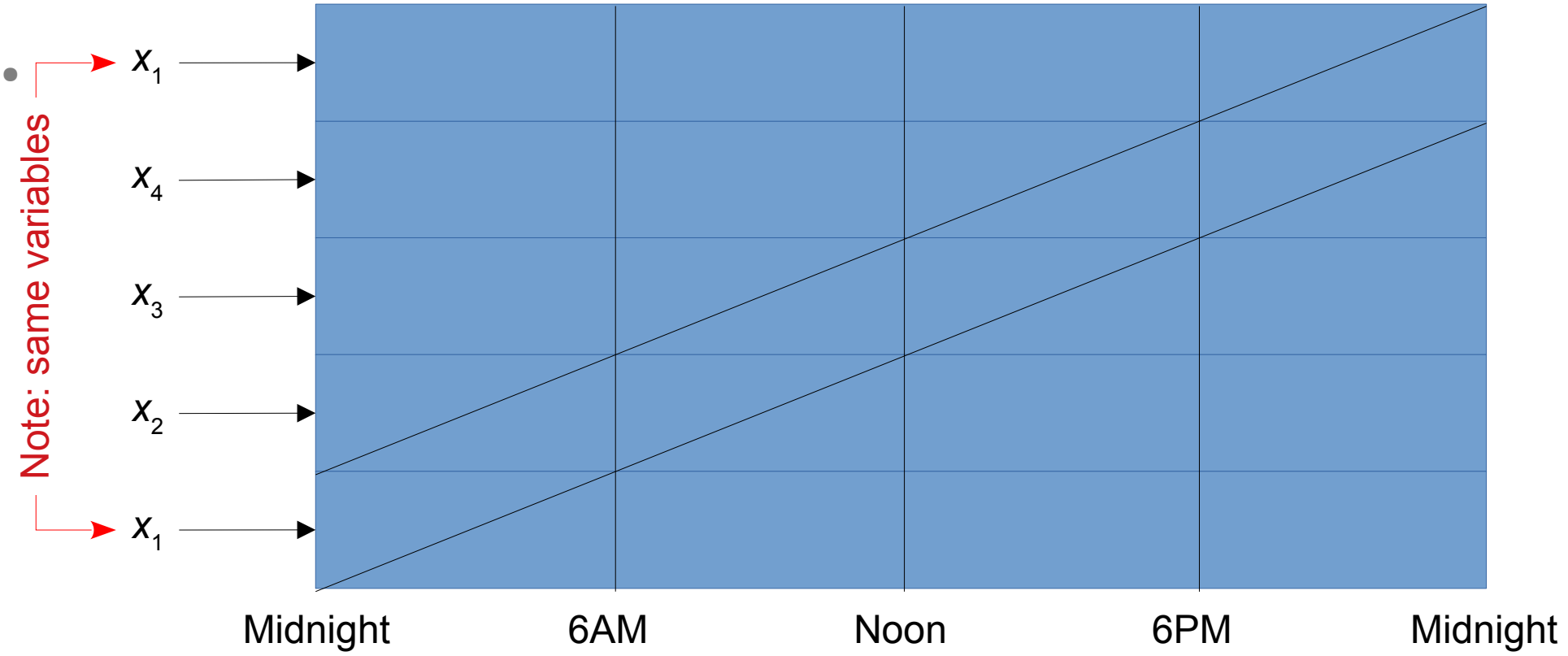
Pabico, 1996.

Interpolation Representation

- Periodic Variables: Values repeat periodically
- Example: Date of year, time of day
- Extreme values are similar
  - December 31 is similar to January 1
  - 11:59 PM is similar to 12:01 AM
- We can capture the similarity by modifying the Interpolation Representation
- Wrap around the representation unto itself

Pabico, 1996.

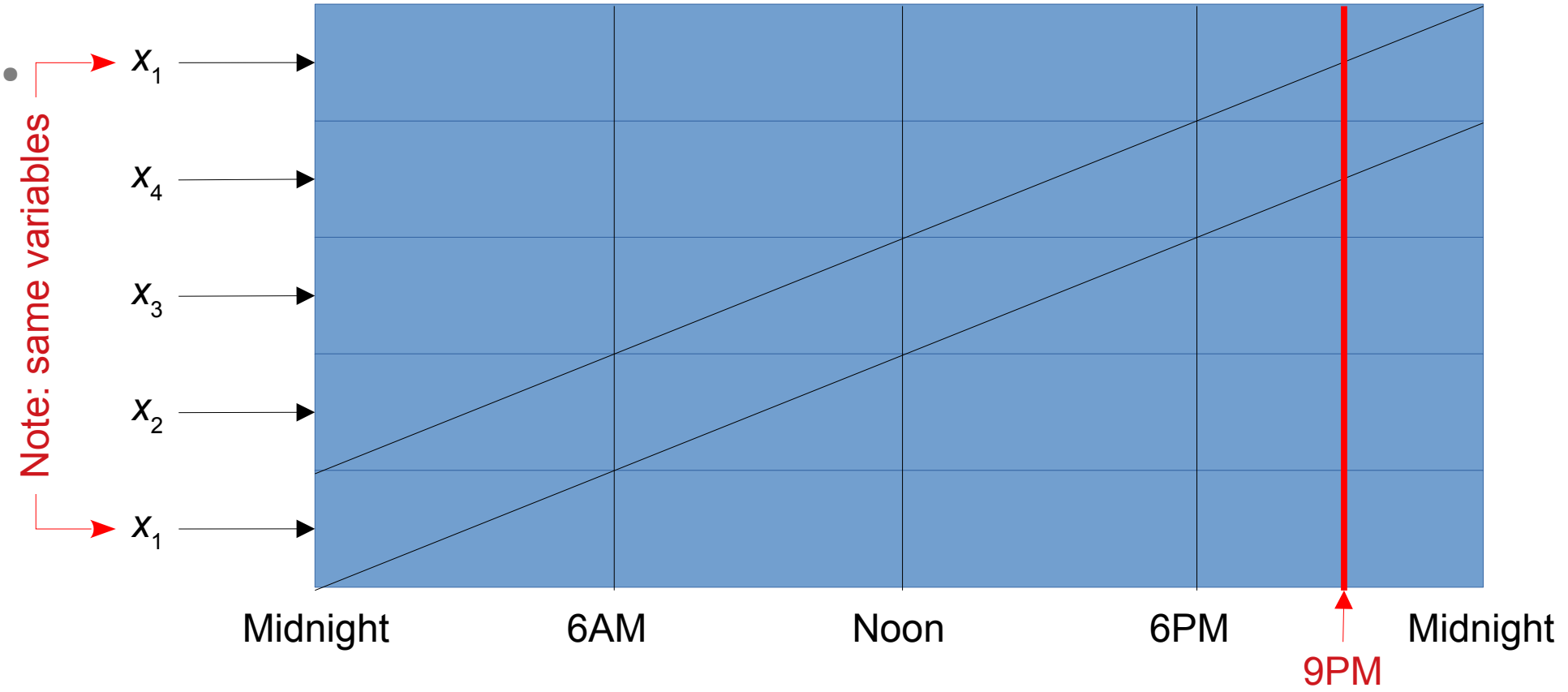
Interpolation Representation



X_1 = 1 at midnight, x_2 at 6AM, x_3 at noon and x_4 at 6PM

Pabico, 1996.

Interpolation Representation



At 9PM, x_1 and x_4 have value = 0.5.

Pabico, 1996.

University of the Philippines
LOS BAÑOS

NEXT ...

Your Puzzling **Questions**
(*and honest, insightful, transformative, advanced comments*)

AND

My (hopefully) Logical & Satisfying **Answers**
(*or nervous blabbers & deflections*)