Automatic Identification Of Ensembles Of Critical Futures in Large Datasets

Amal Sarfraz*[1,2,4], Charles Rougé[1], Lyudmila Mihaylova[1], Jonathan Lamontagne[2], Abigail Birnbaum[2] & Flannery Dolan[3]

*asarfraz1@sheffield.ac.uk, [1]The University of Sheffield, UK, [2]Tufts University, US, [3]RAND Corporation, US, [4]National University of Sciences and Technology, Pakistan

1. Motivation

- Climate integrated assessment models generate large ensembles of future scenarios, revealing the interplay between climate change and socio-economic factors. Our focus is to identify **"Outlier sets"** representing divergent patterns.
- Our analysis employs a subset from Dolan et al. (2021), targeting "Outlier sets" in the 3,000 scenarios for cotton irrigation in the Indus River Basin (IRB) produced by the Global Change Analysis Model, as highlighted in Fig 1.

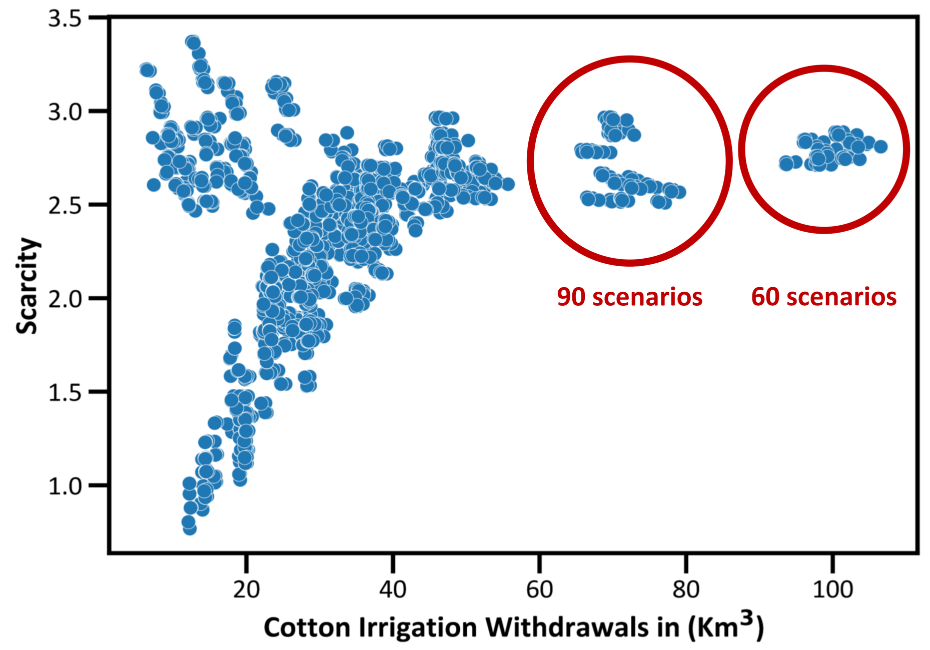

Fig 1. Outlier Sets For IRB Cotton, 2100

2. Methodology

Our novel methodology, Outlier Set two-step Identification (OSTI), involves

I. Identification of candidate outlier sets with Gaussian Mixture Models (GMM) to extract cluster weight.

$$p(x) = \sum_{k=1}^{K} \pi_k \mathrm{N}(x|\mu_k, \Sigma_k)$$

- π_k : mixing weights
- K: number of mixtures (or clusters)
- $\mathrm{N}(x|\mu_k, \Sigma_k)$: Gaussian distribution with x data point, mean μ_k and covariance Σ_k

II. Testing of candidate outlier sets using Inter-cluster Mahalanobis distance (IMD) based statistical tests.

$$IMD^2(\mu_{cl}; \mu) = (\mu_{cl} - \mu)^T * \Sigma^{-1} * (\mu_{cl} - \mu)$$

- μ_{cl}: cluster mean
- μ: dataset's mean
- T: transpose
- Σ^{-1}: inverse covariance matrix

3. Validation

We generated **8000 synthetic datasets** to evaluate OSTI's performance, varying four inlier shapes (circle, ellipse, triangle, irregular) and three outlier parameters (distance (d), angle (θ), standard deviation (σ)) across two cases (one outlier set and two outlier sets) as illustrated in Fig. 2. We assessed results using:

- **F1 score:** It evaluates the ability to find true positives while avoiding false positives and is measured on a scale of 0 to1.
- **Purity**: It is the measure of how much identified sets match synthetic outlier sets on a point-by-point basis.

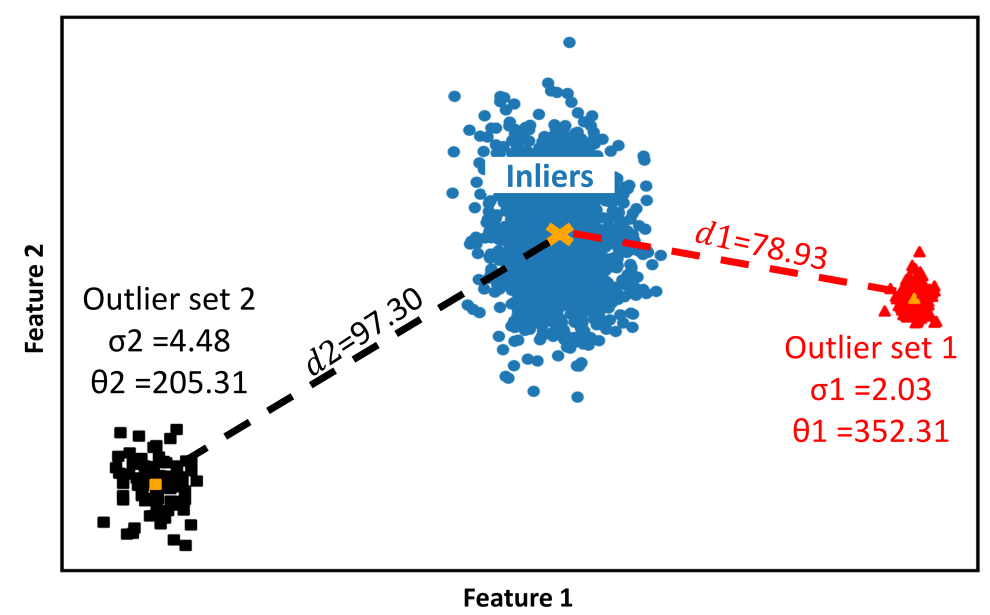

Fig 2 Synthetic Dataset Generation By Varying 4 Inlier Shapes And 3 Outlier Parameters

Table 1. Summary Of Evaluation Metrics For Case 1 And Case 2

Inlier Shapes	Case 1 One outlier set		Case 2 Two outlier sets	
	F1-score	Purity (%)	F1-score	Purity (%)
Circle	0.93	99.98	0.92	99.82
Ellipse	0.88	99.98	0.91	99.82
Triangle	0.93	99.78	0.93	99.29
Irregular	0.94	99.81	0.92	99.22

4. Application

In Fig 3, we highlight major crops in the IRB with all scenarios vs 60 outlying scenarios, illustrating distinct trajectories over time.

- **Early-century** trends show rising withdrawals due to higher summer irrigation needs, extending beyond the traditional cropping calendar.
- **Late-century patterns** reveal increased variability and extreme seasonal irrigation, driven by faster snow and glacier melt.

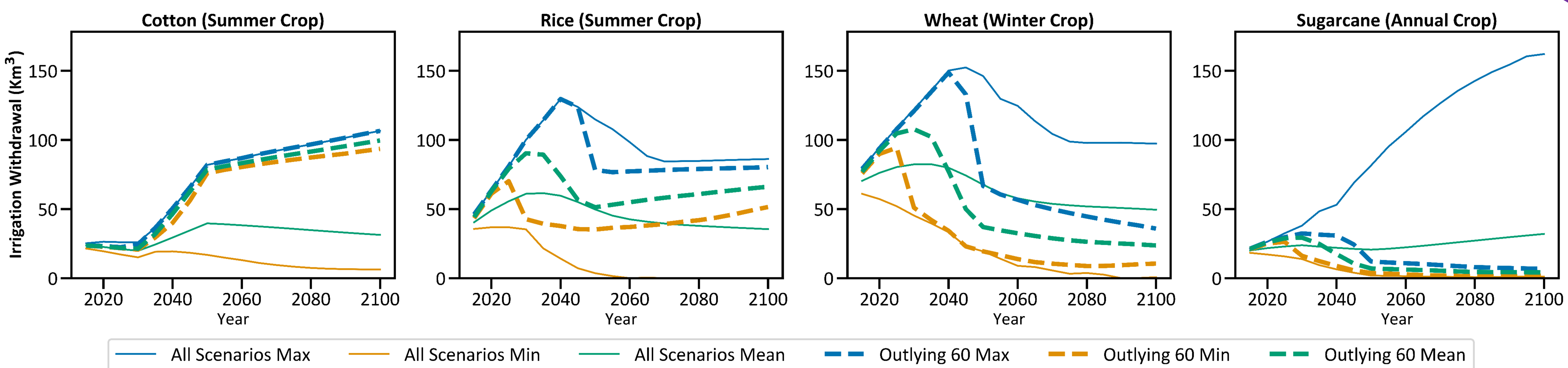

Fig 3. Irrigation Water Withdrawal Trajectories For Major Crops In The Indus River Basin, 2015-2100: All 3000 Scenarios Vs. 60 Outlying Scenarios

References

- Dolan, F., Lamontagne, J., Link, R., Hejazi, M., Reed, P., & Edmonds, J. (2021). Evaluating the economic impact of water scarcity in a changing world. Nat Commun, 12(1), 1915. https://doi.org/10.1038/s41467-021-22194-0
- Calvin, K., Bond-Lamberty, B., Clarke, L., Edmonds, J., Eom, J., Hartin, C., Kim, S., Kyle, P., Link, R., Moss, R., McJeon, H., Patel, P., Smith, S., Waldhoff, S., & Wise, M. (2017). The SSP4: A world of deepening inequality. Global Environmental Change, 42, 284-296. https://doi.org/10.1016/j.gloenvcha.2016.06.010
- Tran, K. C. (1998). Estimating mixtures of normal distributions via empirical characteristic function. Econometric Reviews, 17(2), 167–183. https://doi.org/10.1080/07474939808800410
- P. Mahalanobis, "On tests and measures of group divergence. 1. theoretical formulae," Jour and Proceedings Asiatic Society Bengal, vol. 26, no. 4, pp. 541–588, 1933

COMMONWEALTH SCHOLARSHIPS
Grantham Centre for Sustainable Futures
University of Sheffield