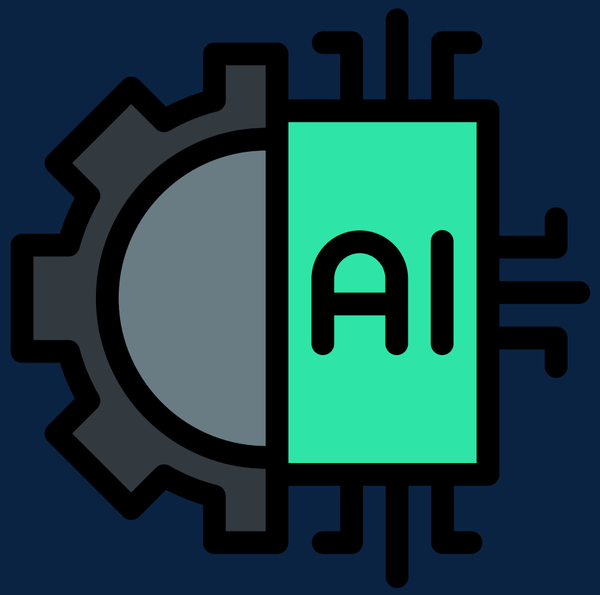
Informative AI Systems

ANWESHAN 2025



PRESENTED BY
PESHWAS

GitHub Repo Link:
https://github.com/archislegend100/Informative_AI_Systems_Peshwas/tree/main_

Problem Statement

- To develop simplified governing equations for cyclone dynamics using the Informative Neural Ensemble Kalman Learning (INEKL) framework.
- Using data from reanalysis or synthetic cyclone trajectories, your system should learn or rediscover relationships governing intensity, moisture, and steering dynamics observed in tropical cyclones.

Input Specifications

- Datasets ERA5

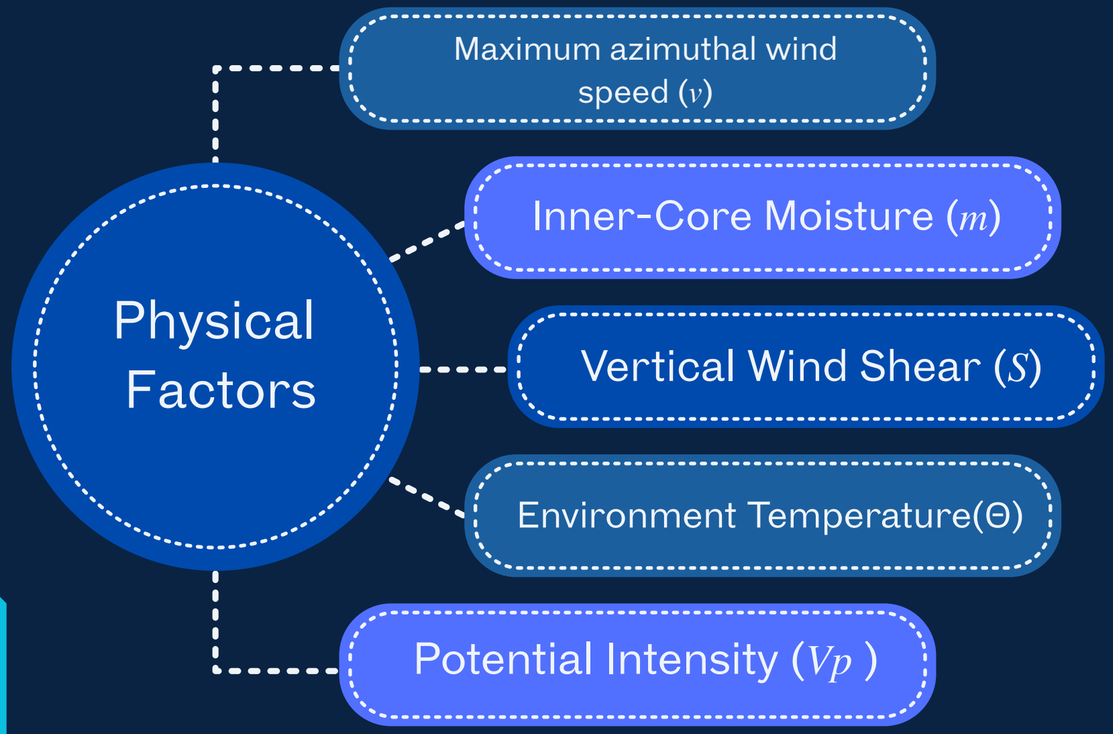
1. Single Levels
2. Pressure Levels

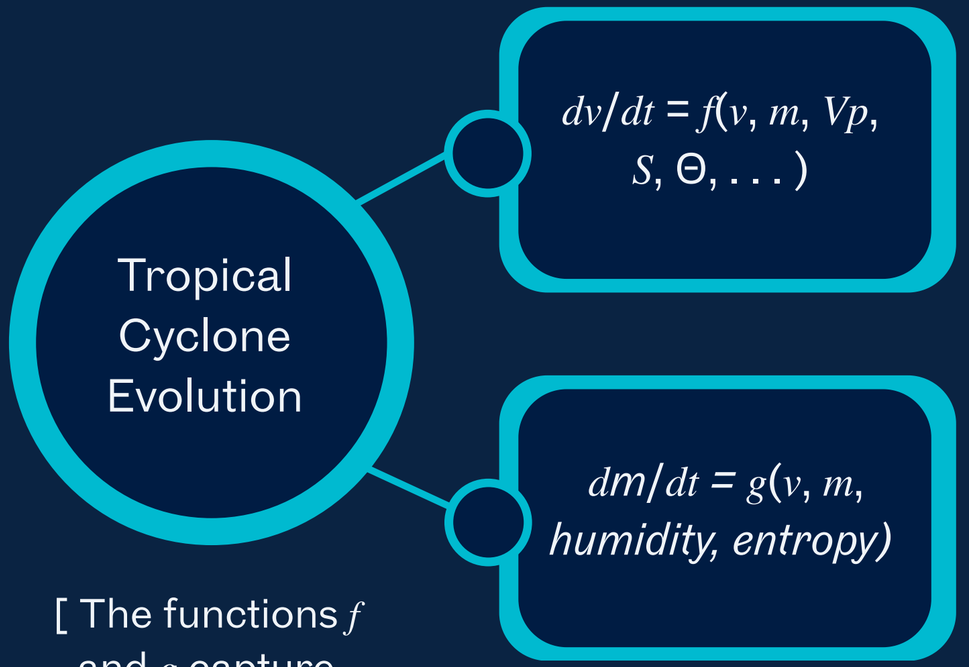
[Inspired from Lin.et.al (2023)]

Output Specifications

- Implement Ensemble Kalman-based learning
- Derive differential equations
- Quantify uncertainty
- Demonstrate information gain

Physics of Tropical Cyclone





[The functions f and g capture intensification and moisture dynamics]

Informative Neural Ensemble Kalman Learning

Ensemble Kalman Update

$$\theta_{k+1}^{(i)} = \theta_k^{(i)} + K_k(y_k - \hat{y}_k^{(i)})$$

where,

$\theta^{(i)}_k$: The state vector for the ensemble member

$\theta^{(i)}_{k+1}$: The updated, corrected state vector

K_k: The Kalman Gain matrix

y_k: The actual measurement.

$\hat{y}^{(i)}_k$: The model's prediction of the observation

Advantages

- Uncertainty Quantification
- Robustness to Noisy Data
- Information-Maximising Updates
- Flexibility in Model Form
- Interpreting equations from data

Dataset Preparation

We have used the following datasets to train and implement our ensemble model:

ERA5 Dataset — To collect atmospheric data(monthly and daily average)

ORAS5 Dataset — To collect the atmospheric boundery layer depth(h)

IBTrACS Dataset — To collect storm paths and wind speed.

After collocating data from all three sources into one csv file, we prepared the derived variables as described in the FAST model.

Apart from this, we also generated synthetic hourly data with FAST baseline model as well to train our model pipeline.

```
RangeIndex: 9367 entries, 0 to 9366
Data columns (total 13 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   storm_id  9367 non-null   int64
 1   t_hr      9367 non-null   float64
 2   v         9367 non-null   float64
 3   m         9367 non-null   float64
 4   Vp        9367 non-null   float64
 5   S         9367 non-null   float64
 6   chi       9367 non-null   float64
 7   alpha     9367 non-null   float64
 8   epsilon   9367 non-null   float64
 9   kappa     9367 non-null   float64
 10  beta      9367 non-null   float64
 11  gamma     9367 non-null   float64
 12  af        9367 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 951.5 KB
```

Model Representation

$$\frac{dv}{dt} = k_1 \alpha \beta v_p^2 m^3 + k_2 v^2 + k_3 \gamma m^3 v^2,$$

$$\frac{dm}{dt} = k_4 v + k_5 m v + k_6 \chi S m.$$

k_1, k_2, k_3, k_4, k_5, k_6 are the weights learned by ensemble.

α, β are the environmental factors derived from datasets.

In FAST learning mode, k_1, k_2, k_3, k_4, k_5, k_6, α, β are set to an arbitrary value Ck/2h = 6×10^{-6}

Model Architecture

Data Preprocessing

- Raw data from ERA5, ORAS5 and IBTrACS are cleaned and processed

Synthetic Data Generation

- A separate, synthetic dataset is created by running the physics-based FAST model.

INEKL Training

- Feature/Target Construction
- Ensemble Initialisation
- INEKL Loop

Simulation

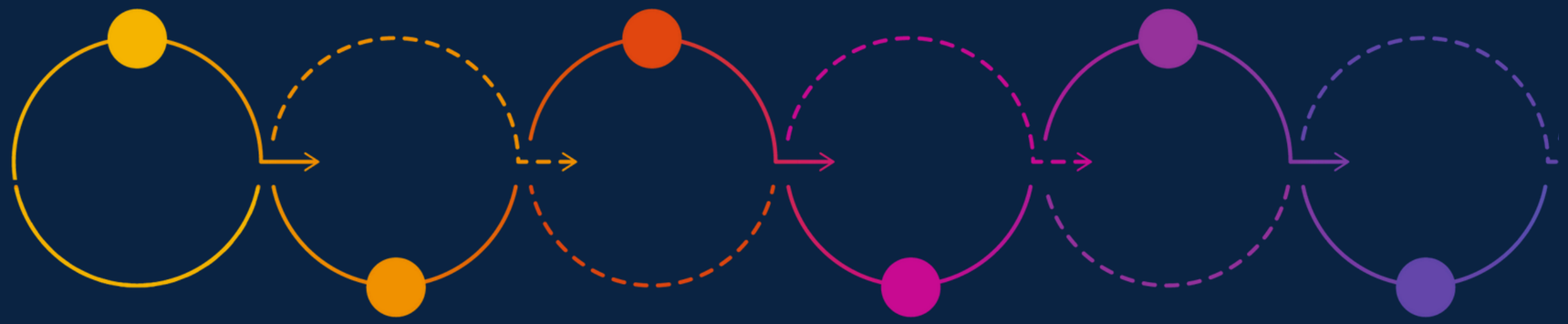
- After INEKL training, we obtain six nonnegative coefficients that mirror the FAST terms.

Data Assimilation

- To bridge the gap between simulation and reality, the model is fine-tuned using real observational data.

Testing and Evaluation

- The final, trained, and assimilated model is evaluated against the FAST baseline model.



Uncertainty Quantification

Uncertainty is quantified by:

- **Ensemble Spread:** Parameter variance across networks directly quantifies prediction confidence.
- **Information Gain:** Mutual information tracking identifies which parameters most reduce prediction error.

Equations:

- Ensemble Spread:

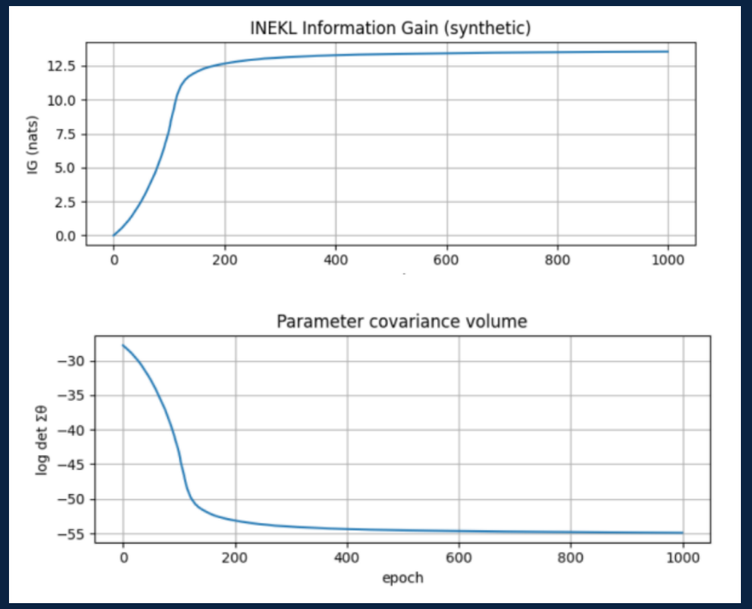
$$\hat{y}_k = (1/N) \sum \hat{y}_k^{(i)}$$

- Information Gain:

$$IG(x^*) \approx (1/2) \log \det(I + H^* P (H^*)^T R^{-1})$$

Data Assimilation

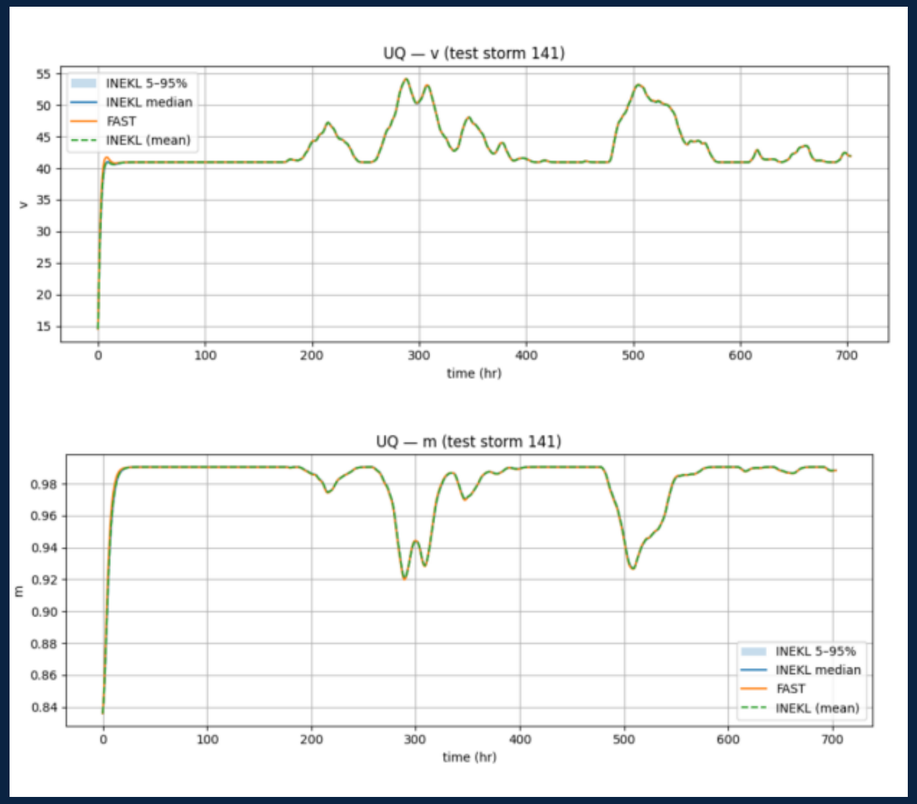
Data assimilation is a technique used to refine ML models using real-world data, allowing it to reduce forecast error. We make use of this in our workflow as well



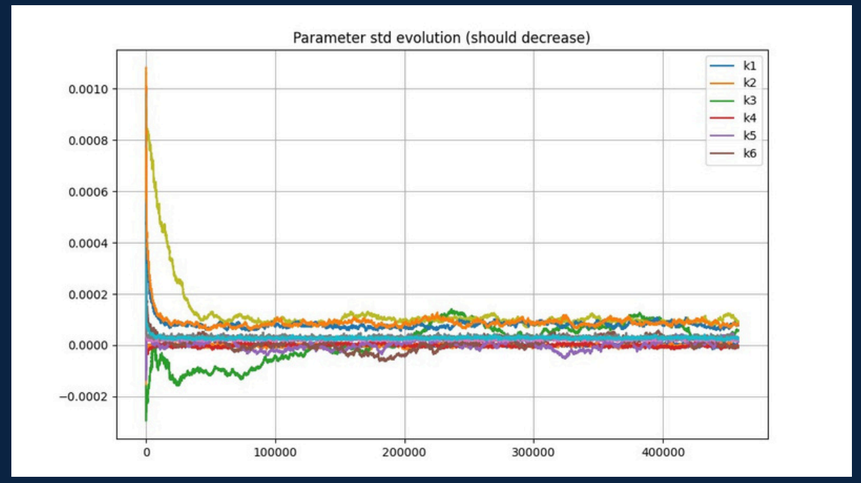
	RMSE_v	R2_v	ACC_v(%)	RMSE_m	R2_m	ACC_m(%)
mean	0.455647	0.976761	98.771991	0.003480	0.989321	98.531048
median	0.409948	0.986627	98.978206	0.002690	0.994860	98.716590
min	0.068627	0.884412	96.981726	0.000527	0.946254	96.852512
max	1.060336	0.999691	99.811488	0.008516	0.999908	99.774860

After first training the INEKL model on synthetic storm data produced using the FAST learning model ODEs, we fine-tune the model on real-world data (ERA5, IBTrACS, ORAS5). This provides much better Mean Squared Error and accuracy

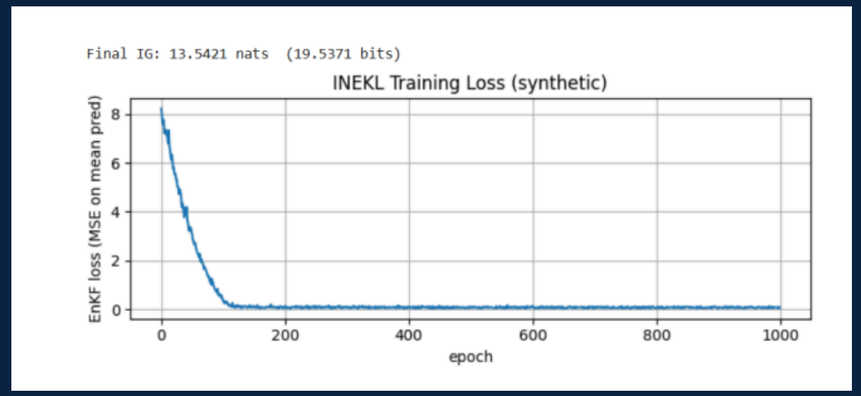
Performance Plots



Uncertainty Quantification



Parameter Evolution



Loss

Results

Quantifiable Learning Progress

- **Information Gain (IG):** The model rapidly assimilates information from early data batches, shown by a sharp rise in IG, before plateauing as parameter uncertainty collapses.
- **Final IG:** 19.57 bits, indicating a significant reduction in parameter uncertainty.

Effective Loss Optimization via Kalman Updates

- The Ensemble Kalman update provides a more stable and efficient convergence compared to traditional gradient-based methods, avoiding local minima and hyperparameter sensitivity

Accurate Trajectory Prediction & Uncertainty Quantification

- Tightened Uncertainty Bands: After data assimilation, the 5-95% prediction ribbons are narrower during stable phases (higher confidence) and widen appropriately during rapid intensification (capturing uncertainty).

Accuracy Comparison

	RMSE_v	R2_v	ACC_v(%)	RMSE_m	R2_m	ACC_m(%)
mean	0.460638	0.976140	87.224699	0.003523	0.989006	91.324384
median	0.414582	0.986318	88.314535	0.002667	0.995102	93.016338
min	0.073305	0.881705	65.605949	0.000508	0.944210	76.380010
max	1.072679	0.999695	98.254703	0.008676	0.999881	98.911120

Figure 5: Results without Data Assimilation

	RMSE_v	R2_v	ACC_v(%)	RMSE_m	R2_m	ACC_m(%)
mean	0.455647	0.976761	98.771991	0.003480	0.989321	98.531048
median	0.409948	0.986627	98.978206	0.002690	0.994860	98.716590
min	0.068627	0.884412	96.981726	0.000527	0.946254	96.852512
max	1.060336	0.999691	99.811488	0.008516	0.999908	99.774860

Figure 6: Results after Data Assimilation

Challenges Faced

- Initial approach of **only training on observational data** lead to high loss, so we approached with data assimilation which increased our accuracy quite a lot.
- Compiling extremely **large datasets**, causing problems in loading to RAM and downloading. Also, these datasets were gated and limited in access
- **Poorly organised definitions** in paper which lead to repeated attempts at creating derived variables. All the parameters are locally dependent and highly variable.

Conclusion

Data assimilation using real observations significantly improved accuracy, reducing RMSE and enhancing calibration.

Developed a hybrid physics–AI framework combining FAST's interpretable ODEs with Informative Neural Ensemble Kalman Learning (INEKL).

INEKL-FAST Hybrid

Information gain analysis validated the learning process and demonstrated efficient knowledge extraction.

Learned storm- and regime-adaptive coefficients, enabling dynamic modeling of tropical cyclone intensity and moisture.

Uncertainty quantification via ensemble variance offered actionable confidence bounds and insight into model reliability.

Thank You