Project G4: Next-Day Wildfire Spread Prediction on mNDWS

- **Milestone II status**: Excellent progress on data engineering, baseline modeling, and careful instrumentation (metrics, calibration, compute footprint). For the remaining milestones, the priority is to (i) *finish and freeze* at least one spatial model (U-Net), (ii) run a small but coherent ablation suite shared across logistic regression and U-Net, and (iii) shape the final report around a small number of clear, quantitative comparisons rather than adding more models.

Core questions for the final report

- **Clarify and tighten your research questions**. You already have good RQs (overall performance, feature families, robustness slices, calibration). For the final report, highlight 2–3 questions that your experiments will *actually answer*, for example:

  - How much does **spatial context** (compact U-Net) improve next-day spread prediction compared with the completed per-pixel logistic regression baseline, in terms of F1/AP and compute cost?
  - Which **feature families** (wind, fuels, vegetation/drought, topography) drive performance for each model, and does the U-Net use them differently than logistic regression?
  - How robust are both models in **challenging conditions** (e.g., high wind, WUI exposure), and does the spatial model reduce failures relative to the linear baseline?

  Try to phrase at least one hypothesis explicitly, e.g.: *"We hypothesize that a compact U-Net improves F1 by 5–10% over logistic regression at similar or slightly higher compute cost."*

- **Define the main comparison table**. Decide now what Table 1 of your final report will be (for example: logistic regression vs U-Net vs simple CNN, with F1, AP, IoU, latency). All remaining experiments and plots should feed into that table.

Baselines and models to finalize

- **Baseline: Logistic regression (already strong)**.

  - Keep the current configuration as your *frozen reference*: channel-aware 1×1 conv, class-balanced BCEWithLogits, fixed pos_weight, calibrated threshold from validation PR curve.
  - Do not continue to heavily tune this baseline; instead, treat its numbers as the anchor against which you compare all spatial models.

- **Model 2: Compact U-Net with ResNet-18 encoder**.

  - Promote the U-Net currently in `mNDWS_UNetModel.ipynb` / `mNDWS_ResNet18UNet.ipynb` into a stable, documented model:
    - Encoder: ResNet-18 (ImageNet pretrained, unless you show it does not help).
    - Decoder: 4 upsampling blocks with skip connections, bilinear upsampling + convolutions.
    - Inputs: a fixed set of channels (e.g., previous burn, wind, NDVI, fuels, topography, drought). Justify which channels you include in the main model.
    - Output: 1-channel sigmoid map, loss = weighted BCE + Dice or focal + Dice (pick one as the default).
  - Run at least one full training run to convergence with fixed hyperparameters and log all metrics parallel to the logistic baseline (F1/AP/IoU, epoch time, tile latency).

- **Optional Model 3: Lightweight CNN or patch-based classifier**.

  - Only add this if you have time. The goal would be to show that U-Net is meaningfully better than a simpler spatial baseline (e.g., small FCN or patch-based ResNet classifier).
  - If you include it, keep architecture description and experiments minimal and aligned with the same metrics and splits.

Metrics, splits, and evaluation

- **Dataset splits and temporal leakage**:
  - You already discuss the imbalance between many pixels and relatively few fire episodes (e.g., ∼185 unique fires in the train split). In the final report, make it explicit how your splits mitigate temporal leakage (same fire across days in different splits).
  - If the mNDWS metadata supports fire IDs or temporal grouping, describe how you use that; if not, acknowledge this as a limitation.
- **Metrics to standardize across models**:
  - Use the same set of metrics for every model: micro F1, IoU, precision, recall, AP, plus calibrated threshold.
  - For each model, report:

    $$\text{F1},\ \text{AP},\ \text{precision at high recall},\ \text{inference latency per tile},\ \text{parameter count}.$$

- **Threshold calibration**:
  - Keep the current strategy: sweep thresholds on validation PR curve, choose the F1-optimal threshold, and then apply it to the held-out test set.
  - Optionally, you can mention future extensions like per-slice thresholds (e.g., different operating points for high-wind vs low-wind), but do not overcomplicate the current project.

Ablation study plan

- **Ablation 1: Feature families (shared across models)**.
  - For both logistic regression and U-Net, run a small set of feature-family ablations:

    $$\text{full model},\ \text{minus wind},\ \text{minus fuels},\ \text{minus vegetation/drought},\ \text{minus topography}.$$

  - Report ΔF1 and ΔAP for each removal. This directly answers "Which feature families contribute most?" for both linear and spatial models.
- **Ablation 2: Dependence on the previous burn mask**.
  - You already show that the previous-day burn mask alone reaches F1 comparable to the full logistic model.
  - For the final report, include:
    - ∗ Logistic regression with and without previous burn mask.
    - ∗ U-Net with and without previous burn mask (same hyperparameters).
  - This demonstrates how much each model relies on persistence vs genuinely forecasting forward spread.
- **Ablation 3: Robustness slices**.
  - For both models, compute metrics on:
    - ∗ high-wind tiles (e.g., wind_75 above the 75th percentile),
    - ∗ tiles with high impervious fraction (WUI-like),
    - ∗ optionally, a few coarse ecoregion bands.
  - Report F1/AP on each slice. Even a small table here will make the final report feel much more aligned with operational questions.

Error analysis and qualitative results

- **Compare qualitative behavior of logistic regression vs U-Net**.
  - Select 3–5 representative tiles:
    - one where both models do well,
    - one where logistic regression mostly repeats yesterday's perimeter but U-Net captures some forward spread,
    - one failure case for both models.
  - Show previous-day burn, model prediction, and next-day ground truth side by side.
- **Summarize systematic errors**.
  - Use confusion matrices or simple statistics to identify common failure modes (e.g., underprediction of new lobes, overprediction at the perimeter).
  - Discuss how these differ between logistic regression and U-Net and whether spatial context reduces any specific type of error.

Code quality and reproducibility

- **Demote notebooks, promote scripts**.
  - Your repo currently relies heavily on notebooks (`mNDWS_LogRegModel.ipynb`, `mNDWS_UNetModel.ipynb`, etc.). For the final report, create:
    - a Python module (e.g., `models.py`) containing the logistic and U-Net architectures,
    - a training script (e.g., `train_logreg.py`, `train_unet.py`) that can be run from the command line with arguments or a config file,
    - an evaluation script (e.g., `eval_models.py`) that reproduces your main tables.
  - Notebooks can remain for exploratory analysis but should not be the only way to re-run experiments.
- **Configuration and environment**.
  - Keep hyperparameters in a config file (YAML or JSON) or a clear `argparse` interface (learning rate, batch size, channels used, loss choice, seeds).
  - Include a `requirements.txt` or `environment.yml` with pinned versions for key libraries (PyTorch, NumPy, etc.).
- **Results logging**.
  - Continue logging loss curves, PR curves, and threshold sweeps, and save the metric summaries to CSV in a `results/` directory with clear filenames (e.g., `logreg_full.csv`, `unet_no_wind.csv`).
  - Ensure that every number in your final tables can be traced back to a specific run and configuration.

Writing the final report

- **Structure**:
  - Introduction: short motivation + clearly stated RQs.
  - Data and Task: concise description of mNDWS, splits, imbalance, and the prediction target.
  - Methods: logistic regression and U-Net architectures, feature sets, losses, training details.
  - Experiments: main quantitative comparisons, ablation results, and robustness slices.
  - Results and Discussion: interpret what the numbers mean for wildfire spread prediction and when spatial models actually help.

  - Limitations and Future Directions: temporal leakage, limited number of unique fires, potential for richer temporal context, etc.

- **Clarity and focus**:
  - Emphasize 2–3 main findings, for example:
    - compact U-Net vs logistic regression (performance and compute),
    - feature-family importance (which covariate groups actually matter),
    - behavior in high-wind / WUI conditions.
  - Avoid spreading effort across many small variants. A clean, well-explained comparison between one strong baseline and one strong spatial model will make the final report much more convincing.