**1. Should I use one model for all viruses or separate models per virus?**

**Assumption:**

You assume that the viruses (Flu A/B, RSV, SARS-CoV-2) share enough underlying dynamics to justify a single model—or at least a shared architecture.

**Assumptions to Examine:**

* **Homogeneity of dynamics**: Do the viruses share transmission patterns, seasonality, lags in wastewater signal vs. infection?
* **Shared features**: Is the same set of input features equally predictive across viruses?
* **Model capacity**: Can one model disentangle multiple virus patterns without confusion?

**Counterpoints:**

* The viruses have **different seasonalities, shedding behaviors, and transmission profiles** (e.g., RSV vs. SARS-CoV-2).
* A shared model risks **information leakage**, or worse, modeling noise rather than signal.
* Separate models may allow for **optimized hyperparameters and architecture** tuned for each virus.

**Alternative Framing:**

* Use **multi-task learning** (e.g., shared base layers, virus-specific output heads).
* Frame as a **multi-output regression/classification** problem (i.e., predicting virus states jointly but not identically).
* Use **shared representation learning** (autoencoders, embeddings) followed by virus-specific decoders.

**Reasoning Test:**

Does one model actually generalize across all viruses better than specialized ones? You can empirically test this via performance benchmarks.

**Recommendation**: Prototype both. Start with a shared model for simplicity but benchmark against separate models. Consider hybrid solutions (e.g., shared encoder + virus-specific heads).

**2. How can I compare and validate the model performance fit for scientific publication?**

**Assumption:**

You assume standard ML metrics (MAE, RMSE, F1-score) suffice.

**Hidden Assumptions:**

* That metrics like F1-score or RMSE capture what’s epidemiologically meaningful.
* That validation across time (non-iid) is appropriately handled.

**Counterpoints:**

* Time-series cross-validation is not the same as k-fold CV. Leakage is a real risk.
* Purely statistical performance (e.g., MAE) may not capture **public health relevance** (e.g., timely detection of *epidemic start*).

**Alternative Framing:**

* Use **epidemic-specific metrics**: correct detection of start week, lead-time to peak, false alerts.
* Define and report **lead/lag error**: how early/late the model detects onset.

**Recommendation**: Use **time-aware validation** (e.g., walk-forward validation), and supplement standard metrics with **epidemic signal-based performance** (e.g., precision in detecting start of n-consecutive positive weeks).

**3. Should I use national level data or WWTP-level data—or both?**

**Assumption:**

You assume national and WWTP-level data can be synergistic.

**Hidden Assumptions:**

* That the noise in WWTP-level data won’t swamp the signal.
* That models can generalize across heterogeneous WWTPs.

**Counterpoints:**

* WWTP data can be **spatially variable**, with inconsistencies in population, flow, and signal detection.
* Aggregation to national level may **smooth out epidemic signals**, hurting early detection.

**Alternative Framing:**

* Use **hierarchical models** (e.g., national-level trend + WWTP-specific residuals).
* Learn WWTP-level embeddings and **predict epidemic likelihood at multiple levels**.

**Recommendation**: Use both. Start with WWTP-level models, and **compare aggregation vs. disaggregation** for model performance. Consider **spatio-temporal modeling** if the goal is localized epidemic detection.

**4. Is 2–4 years of weekly data enough? Should I augment?**

**Assumption:**

You assume data quantity is a limiting factor.

**Hidden Assumptions:**

* That you don’t already have enough signal in 2–4 years of weekly data.
* That more data is always better.

**Counterpoints:**

* Weekly data for 3 years = ~150 time steps per virus. If your model is deep, this is **extremely limited** for supervised learning.
* Data augmentation for time-series (especially epidemiology) is **non-trivial**: you risk introducing biologically implausible scenarios.

**Alternatives:**

* **Synthetic epidemics** using domain-informed rules (epidemic curves, SIR-based patterns).
* **Noise injection**, bootstrapping, or frequency-domain transforms (e.g., Fourier warping).
* Use **unsupervised pretraining** (e.g., masked time modeling, transformers) before supervised fine-tuning.

**Recommendation**: Your data is *marginally sufficient* for shallow models but *weak for deep learning*. Consider **transfer learning or domain-informed augmentation** for robust models.

**5. Should I train incrementally (e.g., national → WWTP) or in bulk?**

**Assumption:**

You assume incremental training offers learning benefits.

**Hidden Assumptions:**

* That model can effectively transfer from one scale to another.
* That bulk training would obscure nuances.

**Counterpoints:**

* Incremental training may lead to **catastrophic forgetting**, unless using continual learning methods.
* Model might **overfit national trends**, then underperform on noisier WWTPs.

**Alternative:**

* Train **joint models** with explicit multi-resolution inputs (WWTP-level + national trends).
* Use **curriculum learning**: progressively increasing complexity of data over epochs.

**Recommendation**: Start with bulk training. If performance degrades on WWTP-specific tasks, explore fine-tuning strategies. Use early stopping and monitor divergence.

**6. Do I need to label future data for prediction tasks?**

**Assumption:**

You assume labels are only needed for classification tasks.

**Hidden Assumptions:**

* That regression (RNA prediction) doesn’t need labeled outcomes (spoiler: it does).
* That "future prediction" can be learned without targets.

**Counterpoints:**

* For predicting next RNA count, you need labeled historical RNA counts (supervised learning).
* For classifying *future* epidemic state, yes, you need labels—ideally based on your n-consecutive positive rule.

**Recommendation**: Yes. You need labeled data for **both tasks**:

* For regression: RNA values (past + future).
* For classification: epidemic states based on a consistent rule (e.g., n-consecutive rule).

**7. Should I use complex NN models (LSTM) or simple models (XGBoost)?**

**Assumption:**

You assume neural networks are superior because they’re more powerful.

**Hidden Assumptions:**

* That your dataset is large enough to benefit from NNs.
* That temporal dynamics need recurrent modeling.
* That NN interpretability is sufficient for scientific publication.

**Counterpoints:**

* **XGBoost** often outperforms deep models on tabular + small time-series datasets.
* LSTMs may overfit or fail to generalize unless trained carefully with enough data.
* NNs can be **less interpretable**—a liability in public health or publication settings.

**Alternatives:**

* Try both and report results.
* Use **attention-based models** (transformers) for better explainability.
* Use **shallow models for interpretability** + deep models for raw performance.

**Recommendation**: Don’t default to NNs. Let data quantity and model goals dictate your choice. Prototype **XGBoost vs. LSTM vs. Transformer**, compare performance and interpretability.