**I3T**

# RAPPORT D'ANALYSE ET DE RECOMMANDATION

## POUR LA PRÉDICTION DU NIVEAU D'UN RÉSERVOIR

**Projet : Optimisation de la Gestion de l'Approvisionnement en Eau d'un Réservoir**

Auteure : **Yesmine Srairi**

Date : 3 juillet 2025

# 1.Introduction

Dans ce projet, on cherche à améliorer la gestion d'un réservoir agricole alimenté par deux pompes contrôlées à distance. Aujourd'hui, ces pompes s'arrêtent ou démarrent quand le niveau dépasse certains seuils, mais cette approche reste réactive et ne permet pas d'anticiper les variations à venir.

Pour rendre la gestion plus efficace, on propose d'utiliser des modèles de **Machine Learning** et **d'Analyse de Séries Temporelles (TSA)**. Grâce aux données historiques, on pourra prédire le niveau du réservoir et ajuster le fonctionnement des pompes de façon proactive, évitant ainsi les débordements et les niveaux trop bas.

# 2. Compréhension des Données

Avant d'aborder les modèles prédictifs, il faut bien comprendre la nature et les caractéristiques des données à disposition. Ces données constituent l'historique des niveaux du réservoir mesurés à des instants précis, ainsi que le débit de chaque pompe, et le débit de sortie enregistré aux mêmes moments.

**Exemple de données**

À titre d'illustration, on présente un extrait simplifié des données collectées à une fréquence horaire :

| Horodatage | Niveau (L) | Débit P1 (L/h) | Débit P2 (L/h) | Débit sortie (L/h) |
|---|---|---|---|---|
| 2025-07-01 08 :00 :00 | 5200 | 800 | 700 | 300 |
| 2025-07-01 09 :00 :00 | 5500 | 800 | 700 | 300 |
| 2025-07-01 10 :00 :00 | 5800 | 800 | 0 | 320 |
| 2025-07-01 11 :00 :00 | 5600 | 800 | 0 | 400 |
| 2025-07-01 12 :00 :00 | 5300 | 800 | 700 | 600 |

*Les variables mentionnées ci-dessus correspondent aux informations actuellement connues, mais d'autres variables pourront être identifiées lors de l'exploration des données une fois le jeu de données réel disponible.*

**Caractéristiques des données**

— **Données temporelles** : La dimension temporelle constitue l'aspect le plus important. En effet, le niveau du réservoir à un instant donné dépend fortement des niveaux mesurés précédemment.

— **Variables explicatives** : Les données intègrent l'horodatage, le niveau du réservoir, les débits individuels des pompes , ainsi que le débit de sortie. Ces mesures correspondent directement aux flux physiques entrants et sortants, éléments essentiels pour une prédiction précise.

# 3. Approches générales en modélisation prédictive

Dans le domaine de la modélisation prédictive, on dispose d'un large éventail d'approches méthodologiques. Dans cette section, on va présenter les principales familles de modèles en évaluant leur pertinence pour cette problématique. On expliquera également pourquoi certaines méthodes sont moins adaptées, afin de se concentrer ensuite sur celles qui s'avèrent les plus appropriées à l'étude.

## 3.1 Modèles de Régression Traditionnels

On utilise fréquemment les modèles de régression classiques pour établir une relation, souvent linéaire ou parfois non linéaire, entre des variables explicatives et la variable à prédire, ici le niveau du réservoir.

Parmi ces modèles, on compte la régression linéaire, la régression polynomiale, ainsi que les forêts aléatoires (Random Forests).

**Pourquoi ces modèles sont moins adaptés à notre cas ?**

— **Absence de prise en compte de la dépendance temporelle** : Ces modèles traitent chaque observation de manière indépendante. Or, le niveau actuel du réservoir dépend fortement de ses valeurs passées. Ignorer cette structure temporelle peut conduire à des prédictions peu fiables.

— **Nécessité d'une ingénierie de variables complexe** : Pour être performants sur des données temporelles, ces modèles nécessitent la création manuelle de variables basées sur le temps (par exemple des moyennes ). Ce travail est délicat et peut ne pas capturer l'ensemble des dynamiques temporelles présentes.

**En résumé**, même si ces modèles sont simples à mettre en œuvre, ils restent limités dans leur capacité à modéliser les dépendances temporelles et les interactions complexes propres à notre problématique.

## 3.2 Modèles de Séries Temporelles (TSA)

On utilise les modèles de Séries Temporelles pour analyser des données ordonnées chronologiquement. Ces modèles exploitent la dépendance entre les observations successives, ce qui les rend particulièrement adaptés aux phénomènes évoluant dans le temps.

Parmi les modèles courants, on trouve ARIMA (AutoRegressive Integrated Moving Average) et SARIMA (seasonal ARIMA)

**Pourquoi ces modèles sont pertinents pour ce projet ?**

— **Prise en compte explicite de la dépendance temporelle :** C'est la caractéristique principale des TSA. Ils modélisent directement la relation entre une observation et ses prédécesseurs, ce qui correspond parfaitement à la dynamique du niveau du réservoir.

— **Détection des tendances et des saisonnalités :**Il est probable que des saisonnalités existent, par exemple une augmentation de la consommation durant l'été, ce qui pourrait entraîner des variations récurrentes du niveau du réservoir.

— **Intégration de variables exogènes :** Certains modèles TSA, tels que ARIMAX ou SARIMAX, permettent d'incorporer des variables externes :
  — les débits des pompes
  — le débit de sortie
  ce qui constitue un avantage important pour améliorer la qualité des prédictions.

### 3.2.1 Modèles ARIMA (AutoRegressive Integrated Moving Average)

On considère ARIMA comme l'un des modèles les plus utilisés en analyse de séries temporelles. Il combine trois composantes principales :

— **AR (AutoRégressive)** : La valeur future est expliquée par une combinaison linéaire des valeurs passées.

— **I (Intégrée)** : Cette étape consiste à rendre la série temporelle stationnaire en différenciant les données, c'est-à-dire en éliminant les tendances pour stabiliser ses propriétés statistiques.

— **MA (Moyenne Mobile)** : Le modèle prend en compte aussi les erreurs passées de prédiction pour améliorer la précision.

Un modèle ARIMA est défini par trois paramètres $(p, d, q)$ correspondant respectivement à l'ordre de l'auto-régression, de la différenciation et de la moyenne mobile.

**Formule :**

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

Où :
— $Y_t$ : valeur du niveau du réservoir à prédire au temps $t$ ;
— $\phi_i$ : coefficients auto-régressifs (AR) ;
— $\theta_j$ : coefficients de la moyenne mobile (MA) ;
— $\epsilon_t$ : erreur aléatoire à l'instant $t$ ;
— $p$ : ordre de l'auto-régression ;
— $q$ : ordre de la moyenne mobile.

**En résumé :** On peut dire que le modèle prédit le niveau futur du réservoir en se basant sur le niveau actuel, la tendance générale, les erreurs de prédiction passées.

### 3.2.2 Modèles ARIMAX (AutoRegressive Integrated Moving Average with eXogenous inputs)

Le modèle **ARIMAX** est une extension du modèle ARIMA qui permet d'intégrer une ou plusieurs variables explicatives externes, appelées **variables exogènes**.

Dans cette étude, ces variables sont :

— les débits des deux pompes,
— le débit de sortie,
— voire d'autres facteurs externes si disponibles.

Grâce à ARIMAX, on ne se base pas uniquement sur l'évolution passée du niveau du réservoir, mais aussi sur l'effet direct des actions externes.

**Formule**

$$Y_t = \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \sum_{k=1}^{K} \beta_k X_{k,t} + \epsilon_t$$

**Où :**

— $Y_t$ : valeur du niveau du réservoir à prédire au temps $t$,
— $\phi_i$ : coefficients auto-régressifs ,
— $\theta_j$ : coefficients de la moyenne mobile ,
— $\beta_k$ : coefficients associés aux variables exogènes,
— $X_{k,t}$ : valeur de la $k$-ème variable exogène au temps $t$ (débit d'une pompe...etc),
— $\epsilon_t$ : erreur aléatoire à l'instant $t$,
— $K$ : nombre de variables exogènes.
— $p$ : ordre de l'auto-régression ;
— $q$ : ordre de la moyenne mobile.

### 3.2.3 Lissage Exponentiel (ETS)

Le lissage exponentiel est une méthode qui attribue un poids plus important aux observations récentes qu'aux plus anciennes. Ainsi, les données les plus récentes influencent davantage la prédiction.

Il existe plusieurs variantes de cette technique qui permettent de gérer aussi bien les tendances que les saisonnalités dans les données.

**Exemple :** Pour prédire le niveau demain, on pourrait attribuer, par exemple, 70 % du poids au niveau d'aujourd'hui, 20 % au niveau d'hier, et 10 % au niveau d'avant-hier.

## 3.3. RNN et LSTM

Dans le domaine de l'intelligence artificielle, les réseaux de neurones récurrents (**RNN**, *Recurrent Neural Networks*) constituent une catégorie de modèles conçus pour le traitement des données chronologiques, c'est-à-dire des données organisées sous forme de séquences temporelles. Ces modèles sont particulièrement pertinents dans les situations où **l'état actuel du système dépend des événements passés**, comme c'est le cas pour

ce projet.

Contrairement aux approches classiques, qui considèrent chaque mesure indépendamment, les RNN intègrent une **mémoire interne** leur permettant de conserver une trace des états antérieurs du système, améliorant ainsi la qualité des prévisions.

**Pertinence dans notre contexte :**

— **Prise en compte des dépendances temporelles longues :** Les réseaux RNN, et plus particulièrement leur version avancée appelée **LSTM (Long Short-Term Memory)**, sont capables de modéliser des relations complexes entre des événements espacés dans le temps. Cela est essentiel dans notre contexte, où des variations anciennes du niveau d'eau, des débits des pompes ou du débit de sortie peuvent continuer à influencer l'état actuel du réservoir.

— **Apprentissage automatique des relations complexes :** Ces modèles sont capables d'**apprendre eux-mêmes les interactions pertinentes entre les variables**, sans nécessiter une modélisation explicite par l'utilisateur. Cela inclut notamment les effets combinés des débits d'entrée, du débit de sortie et des fluctuations naturelles du système.

### 3.3.1. Les Réseaux LSTM

Les réseaux **LSTM** constituent une extension des RNN, spécifiquement conçue pour **mémoriser des informations pertinentes sur de longues périodes**.

Les LSTM reposent sur deux éléments fondamentaux :

— Une **cellule mémoire (Cell State)**, qui conserve les informations essentielles tout au long des étapes temporelles.

— Des **portes de régulation (Gates)**, qui contrôlent le flux d'informations dans et hors de la mémoire.
*Ces portes fonctionnent comme des mécanismes de filtrage :*

— **Porte d'oubli (Forget Gate) :** elle décide quelles informations stockées dans la mémoire ne sont plus pertinentes et doivent être supprimées.
*Exemple pratique : lorsque le réservoir a été entièrement vidé, les données antérieures à ce vidage peuvent être écartées, car elles ne sont plus représentatives de la situation actuelle.*

— **Porte d'entrée (Input Gate) :** elle détermine quelles nouvelles informations doivent être ajoutées à la mémoire.
*Exemple : si les pompes viennent d'être activées et que le niveau du réservoir augmente rapidement, cette hausse sera intégrée dans la mémoire pour influencer les prévisions futures.*

— **Porte de sortie (Output Gate) :** elle contrôle quelles informations issues de la mémoire seront utilisées pour produire la sortie à l'instant présent.

*Exemple : la porte peut privilégier les informations les plus récentes, telles que les débits observés au cours de la dernière heure, plutôt que des événements plus anciens.*

**Synthèse du fonctionnement** À chaque instant, le modèle LSTM :

1. Prend en compte les nouvelles données mesurées (*niveau du réservoir, débits des pompes, débit de sortie...*) ;
2. Met à jour la cellule mémoire, en conservant ou en supprimant certaines informations selon leur pertinence ;
3. Produit une sortie, c'est-à-dire une prévision , à partir de l'état actuel de la mémoire.

Cette sortie est ensuite transmise à l'étape temporelle suivante, assurant ainsi une **continuité dans l'analyse des données temporelles**.

**En résumé :** Grâce à cette architecture, les réseaux LSTM offrent une solution robuste et efficace pour **modéliser des systèmes dynamiques complexes**, tels que le niveau du réservoir, où les relations entre les variables évoluent dans le temps et sont souvent non linéaires.

## 4. Choix méthodologique et démarche de mise en œuvre

Dans le cadre de ce projet ,Le système doit non seulement prévoir les variations du niveau d'eau à court terme, mais également anticiper des situations critiques afin d'adapter le fonctionnement des pompes et d'alerter les utilisateurs en cas de besoin.

Après une analyse des différentes approches existantes, ces deux familles de modèles apparaissent comme les plus pertinentes pour répondre aux exigences du projet :

— **Les modèles de séries temporelles classiques (TSA)** :
Ces modèles, comme *ARIMA* ou *SARIMA*, sont largement utilisés pour la prévision de valeurs chronologiques continues. Ils offrent une approche fiable pour des systèmes linéaires et permettent une mise en œuvre relativement rapide, surtout dans des contextes où les dépendances temporelles sont courtes ou moyennes.

— **Les réseaux de neurones LSTM** :
Cette approche issue du machine learning permet de modéliser des relations beaucoup plus complexes et non linéaires, ce qui est souvent le cas dans des systèmes physiques réels où plusieurs facteurs interagissent . De plus, les LSTM sont capables de gérer des dépendances temporelles longues, ce qui les rend plus robustes lorsque l'historique du niveau influence encore les valeurs futures.

Compte tenu des spécificités du système étudié (variations rapides et complexes, arrêt et activation dynamique des pompes, des variables externes), l'approche **LSTM est**

**privilégiée pour son adaptabilité aux relations non linéaires et sa capacité à capturer des dynamiques complexes**. Toutefois, une première analyse basée sur TSA pourra être envisagée pour établir un modèle de référence plus simple.

**Démarche de mise en œuvre prévue :**

1. **Collecte et préparation des données** :
   — Extraction des historiques de fonctionnement des pompes,
   — Extraction des mesures de niveau du réservoir,

2. **Exploration des approches TSA** :
   — Modélisation initiale avec des méthodes telles que ARIMAX ou SARIMAX,
   — Évaluation des performances et des limitations dans la modélisation des comportements non linéaires.

3. **Conception et entraînement d'un modèle LSTM** :
   — Constitution des séquences temporelles nécessaires à l'apprentissage,
   — Entraînement du modèle sur les données historiques,
   — Validation du modèle sur des périodes de test.

4. **Déploiement dans le système existant** :
   — Intégration du modèle dans un environnement connecté à la base de données SQL,
   — Automatisation des alertes en cas de niveau critique,
   — Automatisation du contrôle des pompes selon les prévisions.

**Conclusion** Dans le cadre de ce projet, les approches TSA et LSTM se révèlent complémentaires. L'objectif est d'exploiter d'abord des modèles simples pour une compréhension de base du système, puis de passer à une approche LSTM plus avancée afin d'obtenir des prévisions robustes et fiables, capables de répondre aux exigences opérationnelles du réservoir.

# 5. Outils, bibliothèques et technologies envisagés

La mise en œuvre des modèles prédictifs repose sur un ensemble cohérent d'outils, de bibliothèques et de technologies, sélectionnés pour leur compatibilité avec l'environnement SQL et leurs performances reconnues dans le domaine du traitement des séries temporelles.

## 5.1. Langage de Programmation

Le langage Python sera utilisé pour le traitement des données et la mise en œuvre des modèles prédictifs

## 5.2. Gestion des Données (SQL)

| Outil/Bibliothèque | Fonction | Compatibilité |
|---|---|---|
| pandas | Manipulation des données | Compatible avec SQL |
| mysql-connector-python | Connexion à la base de données | PostgreSQL, MySQL, SQL etc. |

La première étape consistera à extraire les données de la base SQL, pour les transformer en structures manipulables sous Python (*DataFrames*).

## 5.3. Modélisation des Séries Temporelles Classiques (TSA)

| Bibliothèque | Usage principal | Modèles supportés |
|---|---|---|
| statsmodels | Modélisation statistique | ARIMA, SARIMA, SARIMAX, lissage exponentiel... |

La bibliothèque *statsmodels* fournit des outils robustes pour la modélisation des séries chronologiques linéaires, notamment avec ou sans variables exogènes.

## 5.4. Modélisation par Réseaux Neuronaux LSTM

| Bibliothèque | Usage principal |
|---|---|
| TensorFlow / Keras | Modélisation deep learning séquentielle |

*Keras*, en tant qu'API de haut niveau de TensorFlow, simplifie la conception et l'entraînement des réseaux neuronaux récurrents LSTM.

## 5.5. Évaluation des Modèles

| Bibliothèque | Fonctionnalité | Métriques clés |
|---|---|---|
| scikit-learn | Évaluation des performances | MAE, RMSE, Score $R^2$ |

**Métriques utilisées :**

- **MAE (Mean Absolute Error)** : mesure l'erreur moyenne des prédictions.
- **RMSE (Root Mean Squared Error)** : pénalise davantage les grandes erreurs.
- **Score** $R^2$ : indique la proportion de la variance expliquée par le modèle.

## 5.6. Environnement de Développement

| Outil | Rôle |
|---|---|
| Visual Studio Code | Environnement de développement léger, extensible et optimisé pour Python |

## 5.7. Visualisation des Données et des Résultats

| Bibliothèque | Utilité |
|---|---|
| matplotlib, seaborn | Visualisation des données et comparaison des prédictions aux mesures réelles |

Les visualisations sont essentielles pour valider les résultats des modèles et faciliter leur interprétation.

### Résumé de la Chaîne Technologique

| **Étape** | **Outils Principaux** |
|---|---|
| Extraction de données | SQL + pandas |
| Préparation des données | pandas, scikit-learn |
| Modélisation TSA | statsmodels |
| Modélisation LSTM | TensorFlow / Keras |
| Évaluation | scikit-learn |
| Visualisation | matplotlib, seaborn |
| Développement | Visual Studio Code |

## 6. Schéma global de l'architecture technique et du flux de traitement

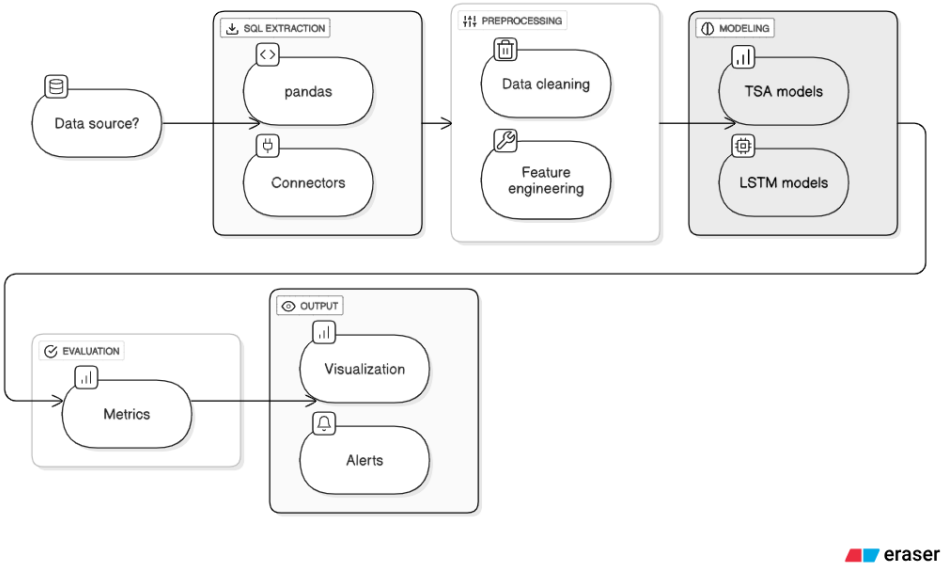


FIGURE 1 – Schéma du flux de traitement des données