Défi EGC 2022 : prévoir l'évolution du niveau de nos nappes phréatiques

Présentation

Le BRGM propose de s'intéresser à la modélisation des niveaux des nappes phréatiques (autrement appelés « niveaux piézométriques » ou « niveau d'eau », ou simplement « niveau »), suivis sous la forme de séries temporelles (aussi nommées « chroniques » dans la suite de ce document) à travers deux défis distincts (il est possible de s'intéresser à l'un, à l'autre, ou aux deux):

1. la prédiction de l'évolution du niveau piézométrique au cours du temps sur une période de trois mois
2. la recherche de motifs ou de séries temporelles représentatives par le biais de méthodes de partitionnement ou de recherche de motifs

Pour chacune de ces deux tâches, des informations externes pourront être utilisées pour contextualiser, enrichir et améliorer les résultats et l'interaction d'un expert avec les données.

Qu'est-ce qu'un niveau piézométrique ?

Contrairement aux eaux de surface (cours d'eau et lacs), les eaux souterraines ne se voient pas mais sont stockées et circulent à des vitesses variées dans le sous-sol, principalement entre les grains de la roche (la matrice) – comme l'eau de mer entre les grains de sable – mais aussi dans les fissures. Les puits et les forages permettent d'atteindre les zones contenant de l'eau (zone saturée en eau), et donc de mesurer la profondeur de l'eau par rapport au sol : c'est le **niveau piézométrique**. Des **capteurs piézométriques** immergés dans des puits, forages ou sources, permettent de donner une image de l'évolution au cours du temps du niveau piézométrique, en mesurant en continu la profondeur de l'eau par rapport au sol[1]. En mesurant le niveau piézométrique en différents endroits, il est possible de donner une image de la circulation de l'eau dans le sous-sol : l'eau s'écoule des niveaux les plus hauts, vers les niveaux les plus bas.

Vous trouverez des informations pédagogiques sur l'hydrogéologie dans la section « Phénomène » de ce document.

Dans ce concours, que vous choisissiez le défi n°1 ou n°2, vous aurez à vous intéresser à l'évolution du niveau de plusieurs nappes phréatiques, aux endroits précis où sont installés les capteurs piézométriques. Ces capteurs mesurent et enregistrent les mouvements des nappes depuis plusieurs années.

Le défi n°1 concerne uniquement une sélection de 18 capteurs piézométriques. Vous trouverez dans

points_eau.csv

le fichier CSV joint certaines données statiques concernant ces 18 piézomètres,

[1] Cf. http://sigesbre.brgm.fr/Qu-est-ce-que-la-piezometrie.html

identifiés par leurs codes[2] et associés à des coordonnées géographiques précises. Ce fichier est encodé en UTF-8. Les colonnes de ce fichier sont décrites dans le tableau ci-dessous.

Nom de colonne	Description
CODE_BSS	Identifiant historique des forages souterrains (encore très utilisé) et - par extension - du piézomètre intégré dedans le cas échéant
BSS_ID	Nouvel identifiant technique du piézomètre, alternatif au code BSS.
LONGITUDE	Longitude du piézomètre
LATITUDE	Latitude du piézomètre
CODE_INSEE_COMMUNE	Code insee de la commune
NOM_COMMUNE	Nom de la commune
CODE_STATION_HYDRO	Identifiant de la station hydrologique[3] éventuellement associée à un piézomètre par des experts hydrogéologues en vue de la modélisation de la chronique piezo
NOM_STATION_HYDRO	Nom associé à la station hydrologique
CODE_BDLISA	Identifiant de la masse d'eau souterraine (aquifère ou nappe phréatique) « surveillé » par le piézomètre. Les masses d'eau sont répertoriées dans la base de données Lisa (BDLisa)
NOM_ENTITE_BDLISA	Nom associé à l'entité BDLisa (masse d'eau).

Ces capteurs piézométriques sont toujours en activité et de nouvelles observations viennent en permanence (au pas de temps quotidien généralement) agrandir la chronique. Toutes les observations sont récupérables notamment grâce à l'API Hub'Eau de piézométrie[4].

Exemple : pour récupérer la chronique, c'est-à-dire l'ensemble des observations stabilisées réalisées par le capteur piézométrique 03124X0088/F[5], depuis sa première mesure (2004) jusqu'à aujourd'hui (5611 observations au moment où nous écrivons), au format JSON, il suffit d'exécuter la requête suivante dans un navigateur ou via un script :

https://hubeau.eaufrance.fr/api/v1/niveaux_nappes/chroniques?code_bss=03124X0088/F

Pour obtenir la même chose au format CSV :

https://hubeau.eaufrance.fr/api/v1/niveaux_nappes/chroniques.csv?code_bss=03124X0088/F

Les champs restitués qui vous intéresseront au premier chef sont *code_bss*, *date_mesure* et *niveau_nappe_eau*.

[2] Deux identifiants possibles pour chaque piézomètre : le « CODE_BSS » (code historique) et le « BSS_ID » (nouveau code)
[3] Une station hydrologique mesure la hauteur et/ou le débit d'un cours d'eau à intervalle régulier (jusqu'à toutes les 5 minutes pour certaines stations, en cas de crue du cours d'eau surveillé).
[4] Il y a deux « endpoints » fournissant les mesures : *chroniques* et *chroniques_tr*, selon que l'on s'intéresse respectivement aux données 'stabilisées' ou aux données 'brutes temps réel'.
[5] Installé dans la commune de Rostrenen en Bretagne, ce capteur fait partie des 18 piézomètres du concours

A noter que pour un capteur donné, la fréquence de ses mesures n'est pas nécessairement constante dans le temps : pour les stations de mesure anciennes en particulier, il n'est pas rare que cette fréquence augmente avec les années.

De quoi dépend la variation du niveau piézométrique au cours du temps ?

Les aquifères (ou nappes phréatiques) sont schématiquement sous l'influence de deux catégories de forçages :

- les flux d'eau entrant dans l'aquifère : le plus généralement : la pluie, modulée des effets liés au ruissellement en surface, de la part consommée par les plantes *via* les racines, etc.)
- les flux d'eau sortant de l'aquifère : le plus généralement par les sources, l'alimentation des rivières/fleuves, les pompages[6].

D'autres données complémentaires pour décrire l'évolution temporelle du niveau des nappes phréatiques peuvent donc être nécessaires. Par exemple des données météo ou hydrologique (voir la section *Données*) permettent d'estimer les quantités d'eau entrantes et sortantes de l'aquifère.


data_Rostrenen.csv

A titre d'exemple, vous trouverez dans ce fichier un ensemble de données compilées en date du 28/02/2018, associé au piézomètre de Rostrenen 03124X0088/F et servant à sa modélisation aux hydrogéologues du BRGM.

Pourquoi modéliser et prédire l'évolution du niveau piézométrique ?

L'eau souterraine contribue à divers usages, dont notamment l'alimentation en eau potable (2/3 des volumes d'eau destinée à l'alimentation humaine proviennent de ressources souterraines), l'irrigation, ou des usages industriels (refroidissement, lavage, etc.), mais participe également à soutenir les débits d'étiages des cours d'eau pendant les périodes sèches.

Pour assurer la pérennité de ces différents usages, ou fonctions, il est nécessaire de comprendre l'évolution du niveau piézométrique au cours du temps, et de pouvoir prédire sont évolution. Cela peut notamment servir à mieux répartir les volumes disponibles pour chacun des usages, orienter et adapter les pratiques : cultures moins gourmandes en eau, désimperméabilisation des sols, etc… et donc plus globalement, cela participe à une gestion responsable d'un territoire.

Objectifs

Dans ce défi, nous vous proposons deux objectifs. Chaque participant peut répondre à l'un ou l'autre de ces objectifs, ou aux deux.

Objectif n°1 : la prédiction de chroniques piézométriques

Prédire l'évolution des 18 chroniques correspondant aux 18 capteurs piézomètres sélectionnés, sur un horizon de 3 mois – plus exactement **à partir du 15 octobre 2021 jusqu'au 15 janvier 2022 inclus** – avec un **pas de temps quotidien**, soit 93 valeurs à estimer par chronique. Ces prévisions sont à

[6] Les piézomètres sélectionnés ne sont pas influencés par des pompages. Ce terme peut donc être négligé a priori.

remettre au plus tard le 15 octobre à 00:00. L'écart entre ces prévisions et la réalité sera évalué à partir du 16 janvier 2022 automatiquement grâce à la métrique de l'erreur moyenne quadratique réduite (Root Mean Squared Scaled Error – **RMSSE**). La métrique est calculée pour chaque série comme suit :

$$\boldsymbol{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \widehat{Y_t}\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}(Y_t - Y_{t-1})^2}}$$

où Y_t est la valeur réelle de la série chronologique examinée à l'instant *t*, $\widehat{Y_t}$ la prévision générée, *n* la longueur de l'échantillon d'apprentissage (nombre d'observations historiques) et *h* l'horizon de prévision.

Le choix de cette métrique se justifie comme suit :

- Elle est indépendante de l'échelle, ce qui signifie qu'elle peut être utilisée efficacement pour comparer les prévisions de séries à différentes échelles.
- Contrairement à d'autres mesures, elle peut être calculée en toute sécurité car elle ne repose pas sur des divisions avec des valeurs qui pourraient être égales ou proches de zéro.
- Elle est symétrique : elle pénalise les erreurs de prévision positives et négatives, aussi bien sur les petites et grandes valeurs prévues, de manière égale.

Le cumul des scores RMSSE des 18 séries chronologiques attendues donnera le score final de chaque participant et permettra de classer les participants entre eux, le plus petit score cumulé étant le meilleur.

Objectif n°2 : le regroupement par famille de comportement

Le second objectif correspond à la question plus ouverte de l'**exploration** et la **représentation** des données.

Dans ce défi, on demande de **considérer l'ensemble des piézomètres accessibles via l'API Hub'eau**, soit plus de 20 000 stations piézométriques et donc autant de chroniques. Certaines de ces chroniques sont très longues et/ou très denses, d'autres beaucoup moins ; certaines présentent peu de données manquantes, d'autres davantage...

L'**exploration des données piézométriques** peut se faire de multiple façon. Une proposition consiste à travailler sur du clustering : regrouper et représenter des données *similaires*, à la fois pour pouvoir représenter la diversité des données des 20 000 stations, et pour exposer les différents forçages qui s'expriment dans les données. Pour les nappes phréatiques, il pourrait s'agir de détecter et représenter différentes familles de comportements, à partir des données de chroniques piézométriques, ou en ajoutant des données complémentaires : pluviométrie, évapotranspiration, type d'aquifère (BDLISA), hydrométrie etc.

Il est par exemple observé que certaines nappes phréatiques réagissent plus rapidement que d'autres aux évènements météo, certaines présentent des cycles annuels, pluriannuels, etc. L'échelle des regroupements, c'est-à-dire le nombre de groupes, est laissée au choix des participants. Différentes échelles peuvent être pertinentes.

D’autres types de caractéristiques peuvent être explorées et représentées : répartitions temporelles des extrêmes, vitesses de vidange (périodes de baisse du niveau d’eau), répétitions de motifs, etc. en fonction de la position géographique, de la profondeur, du type de réservoir…

Les résultats de l’exploration devront être représentés de façon compréhensible pour les experts hydrogéologues du BRGM. La représentation des données, et donc la recherche de solutions de visualisations originales est un objectif important de ce second défi.

La **représentation des données piézométrique** est d’ailleurs un objectif suffisamment important pour éventuellement être l’objet d’étude unique de ce défi n°2. La représentation des données de séries temporelles est habituellement réalisée sous la forme de graphiques binaires, exposant l’évolution du niveau piézométrique au cours du temps mais de nombreuses autres représentations sont possibles, et pourraient faire ressortir des phénomènes cachés dans les données, en exploitant au mieux les aptitudes visuelles avancées de l’espèce humaine.

Le résultat final sera évalué par des experts hydrogéologues du BRGM. Cette évaluation ne donnera pas lieu à un classement mais à des remarques, questions et suggestions, amorçant potentiellement de riches échanges et d’éventuelles collaborations.

Soumission

Pour répondre au **défi n°1**, vous devrez soumettre vos prédictions – avant le 15 octobre 2021 donc – sur le site Easychair du défi 2022 accessible à l'adresse suivante :

https://easychair.org/conferences/?conf=egc2022

Le fichier de résultat devra être un fichier TEXTE de type CSV contenant 3 colonnes nommées (sur la première ligne) : CODE_BSS, DATE, NIVEAU_PIEZO

Exemple :

```
CODE_BSS, DATE, NIVEAU_PIEZO
07548X0009/F, 2021-10-15, 132.5
07548X0009/F, 2021-10-16, 134
07548X0009/F, 2021-10-17, 137
etc…
```

Note : les dates doivent être exprimées au format : AAAA-MM-JJ

Ce fichier devra être nommé tel que : NOM1_NOM2_..._NOMn.csv où NOMx représente les noms des auteurs du travail et de la publication.

Pour répondre au **défi n°2**, vous devrez soumettre votre résultat, quelle que soit sa forme, également avant le 15 octobre 2021 et sur le site Easychair du défi 2022 accessible à l'adresse suivante : https://easychair.org/conferences/?conf=egc2022

Quelle que soit le format du livrable, il doit être lisible facilement avec des outils classiques : images raster (jpeg, png, gif, etc.), ou vecteur (principalement shapefile), notebook Jupyter, etc.

ANNEXES

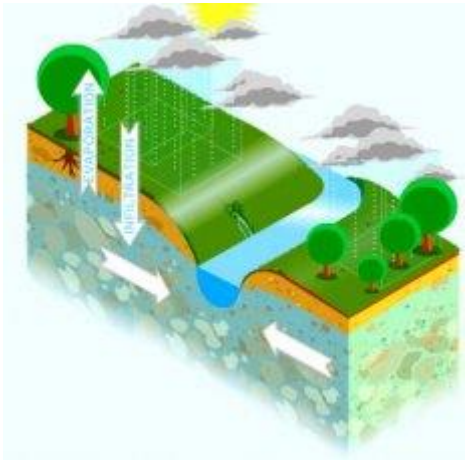
Enjeux

La prévision de niveaux de nappes en période de basses eaux (l'été) peut permettre aux autorités d'anticiper les risques d'éventuelles défaillances des captages servant à l'alimentation en eau potable, comme par exemple dans l'est de France à Hyémondans et Lanthenans (département du Doubs) : plus d'eau au robinet pendant l'été 2020.

La prévision de niveaux de nappes en période de hautes eaux (l'hiver) peut aussi présenter un intérêt dans les zones où la nappe affleure à la surface du sol. Une nappe trop haute peut entrainer une saturation du sol et localement des inondations par remontée de nappe, comme à Sost et Rabastens-de-Bigorre (Hautes-Pyrénées), inondées durant l'été 2018.

Phénomène

Vous trouverez davantage de contenu pédagogique, par exemple sur le site du Système d'Information pour la Gestion des Eaux Souterraines (SIGES) de la région Centre-Val-de-Loire.



L'évolution temporelle du niveau piézométrique – aussi appelé charge hydraulique – dépend schématiquement de l'équilibre entre les entrées et les sorties du système « nappe en sous-sol », soit :

1) La quantité d'eau qui rentre dans le système : les pluies efficaces correspondant à la part de précipitations qui ruisselle à la surface du sol et qui s'infiltre jusqu'à la nappe (le reste étant soit évaporé, soit utilisé par la végétation). En raison de l'élévation des températures et du développement de la végétation, ces pluies efficaces sont faibles d'avril à septembre, et plus importantes d'octobre à mars ; la connexion à d'autres aquifères, les pertes de rivières vers la nappe (en cas de « connexion nappe-rivière » et d'un niveau d'eau de surface supérieur à celui de la nappe) ;

2) La quantité d'eau qui sort du système : par les sources ; les fuites dans les rivières (en cas de « connexion nappe-rivière » et d'un niveau d'eau de surface inférieur à celui de la nappe), ou les prélèvements humains (eau potable, irrigation agricole ou prélèvement industriel).

La vitesse de réaction du système est fonction de divers paramètres tels que sa taille, sa géométrie ou encore la nature géologique du réservoir.

Contexte

Ce défi est proposé par le Bureau de Recherches Géologiques et Minières (**BRGM**).

Actuellement, afin de faire des prévisions du niveau des nappes en des endroits précis, le BRGM utilise deux modèles dits « globaux », appelés Gardénia et Tempo ©BRGM. Ces modèles permettent, à partir des données météorologiques, hydrologiques et piézométriques d'entrée (associées ou non à des données de prélèvement), de simuler un niveau de nappe.

Gardénia est un modèle hydrologique global à réservoirs (modèle physique)[7]. Tempo est un logiciel de caractérisation du fonctionnement hydrogéologique à l'aide de fonctions de transfert (modèle de traitement du signal de type « boite noire »)[8]. Dans les deux cas, une modélisation est entreprise après avoir une bonne connaissance géologique et hydrogéologique du secteur à étudier.

En tant que gestionnaire du réseau piézométrique national dans le cadre de sa mission d'appui aux politiques publiques, le BRGM est particulièrement intéressé par la modélisation des niveaux piézométriques et a en permanence la volonté d'explorer, de comparer et de développer de nouvelles approches.

Ce défi s'inscrit dans cette démarche d'exploration : en ouvrant le sujet à la communauté « data science » française, l'ambition est double : intéresser une partie très dynamique de la communauté scientifique, en vue de découvertes et de futures collaborations, et fournir un « benchmark » de référence sur le domaine.

Données

Les données habituellement utilisées pour modéliser et prédire l'évolution des niveaux de nappes avec des modèles globaux sont :

- Piézométriques : profondeur des eaux souterraines, la variable cible
- Météorologiques : pluie et évapo-transpiration potentielle (ETP) → pour estimer la recharge de la nappe
- Hydrologiques : hauteur et débit des eaux de surface (rivières, lacs) → permet d'estimer la partie des pluies qui ruissellent en surface – à l'intérieur d'un bassin versant – et la partie d'eau de rivière qui provient de la nappe
- Prélèvements : que ce soit pour l'eau potable, l'industrie ou l'irrigation agricole, c'est la partie des données la plus difficile à obtenir

[7] Cf. https://www.brgm.fr/production-scientifique/logiciels-scientifiques/gardenia-logiciel-modelisation-hydrologique
[8] Cf. https://www.brgm.fr/projet/tempo-outil-modelisation-gestion-hydrosystemes

Ces données sont généralement associées à une bonne connaissance géologique et hydrogéologique du secteur à étudier (qui permet un paramétrage fin des modèles Gardénia ou Tempo).

Il existe plusieurs sources de données potentielles pour chacun des aspects (piezo, hydro, météo…), gratuites ou non, plus ou moins facile à utiliser, etc. Nous conseillons ici une liste de sources de données (publiques et gratuites) mais sans contraindre les participants au défi. Chacun peut proposer et exploiter d'autres sources de données s'il le juge pertinent.

- Les données piézométriques Françaises via
  - Le web service Hub'Eau : https://hubeau.eaufrance.fr/page/api-piezometrie

- Les données hydrologiques
  - Accès à toutes les stations hydrométriques publiques
    https://www.data.gouv.fr/fr/datasets/stations-hydrometriques-metropole/
    http://www.geocatalogue.fr/Detail.do?fileIdentifier=13f9ce88-7764-494e-ae8f-2d737b5eaaf8
  - Banque HYDRO
    - Consultation (pas de compte nécessaire) : Rechercher des données, Consultation, Recherche d'une station : par code ou nom de rivière
    - Téléchargement (nécessité de disposer d'un compte)

- Les données météorologiques
  - Cf. portails européens et français vers diverses données de climat, dont :
    - Données européennes du ECMWF : https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5
      - Maille 25km x 25km
    - ETP en grille "Reference evapotranspiration" (Aridity/continentality indices, number 91, ID Eto) - Monthly data extracted from the grid available
    - Données maillées quotidiennes à l'échelle du globe - Résolution : 0,25° (~25 km) : Variables : pluie totale, ETP – REANALYSES ERA5 hourly data on single levels from 1979 to present – Mise à jour quotidiennement - Données disponibles 5 jours avant le temps réel

- Référentiel hydrogéologique : https://bdlisa.eaufrance.fr/

- Contexte géologique : les cartes géologiques de la France au 1/50 000
  - par le site InfoTerre : http://infoterre.brgm.fr/formulaire/telechargement-cartes-geologiques-departementales-150-000-bd-charm-50

Contacts

defi-egc-2022@brgm.fr