Response to reviewer hess-2019-166
The AquiFR hydrometeorological modelling platform as a tool for improving groundwater resource monitoring over France: evaluation over a 60 year period.

Anonymous Referee #2

General comments

Comment: *1 It is not clear if the runoff is calculated with a common code and/or grid (P15L17 suggest that it is not the case), but nothing is explained about how the river routing is performed in the different basins.*

Response: Yes, indeed, some information was lacking. Surface runoff is computed using the land surface scheme of SURFEX on an 8 km resolution grid. This 8 km resolution grid corresponds to the grid provided by the SAFRAN atmospheric analysis. The surface runoff is then routed to the river by the hydrogeological models, with their own spatial resolution (varying from 100 m to 8 km). To be clearer, in the revised manuscript, the section 2 is entirely modified. It includes a new scheme (see Figure 1 of the present document) presenting the physical interaction between the modules and the main processes accounted for the estimation of the water flows, and on a new version of the former Figure 1 (see Figure 2 of the present document) that better presents the technical connection between the module.

The authors agree that no information was given on the river routing. A paragraph is now added in section 2 : "*River routing is performed based on kinematic wave approach in MARTHE and by the RAPID model based on the Muskingum approach (David et al., 2011) in EAUDYSSEE. River-groundwater exchanges are in both directions for all the models. Each regional model uses its own river network at its own resolution*".

Comment: *2. Are the rivers connected bidirectionally with the groundwater? P6L28-29 suggests it can be done, but is it done?*

Response: Yes rivers are connected in both direction in the MARTHE and EauDyssée models. A sentence is added in the text (see answer above). This information is also now provided in section 2.3 (EauDyssée) and 2.4 (MARTHE) that presents the hydrogeological models and is shown in the new Figure 1.

Comment: *3. EROS are lumped models that simulate karst in a simpler way. This is reasonable. It is mentioned that AquiFR will be used for climate change studies, but it is not mentioned how the calibrations of these lumped models will hold in a changing climate.*

Response: A new section "3. Methodology" is added to the manuscript for describing the models, the calibrations and the statistical criterias used for the evaluation. A subsection of this new section "3. Methodology" is now devoted to the calibration of the models. It is now stated that "*For the karst system software EROS, the models were calibrated based on the SAFRAN atmospheric analysis, by using an optimization of the statistical comparison between observed and simulated daily riverflows*." This new section is presented at the end of this document.

It is true that part of the uncertainty of the impact of climate change on the karst systems is linked to the hydrological model and to its calibration. But it is beyond the scope of this article to discuss such uncertainty.

Comment: *4. It is not clear if there is a bidirectional coupling between the aquifer and the soil (SURFEX). P6L15 says it can be done. Figure 1 shows and arrow that goes from the post-processing to SAFRAN/SURFEX, but it is not clear what it means. Is there a bidirectional coupling between soil and aquifer? Is SURFEX just a forcing or at each time step it is updated with information coming from the aquifers?*

Response: Thanks to stress out this important point. Although capillary rise can be accounted for in SURFEX, in the current version of AquiFR, no bidirectional coupling between the aquifer and the soil is taken into account. This is no clearly stated in section 2: "*In this version 1.2 of AquiFR, no feedback from groundwater to the soil of SURFEX is taken into account. Therefore, a preliminary step illustrated by Figure 2**Erreur ! Source du renvoi introuvable.**a is to estimate groundwater recharge and surface runoff with SURFEX taking into account the atmospheric forcing from SAFRAN prior to an OpenPALM run".* In section 2.2 presenting SURFEX, it is now stated: "*In the present study, no bidirectional coupling between the soil of SURFEX and the aquifers is taken into account. Thus, a one-way coupling from the soil of SURFEX to the aquifer is taken into account in order to provide groundwater recharge and surface runoff to the AquiFR platform.* " . The former Figure 1 was modified to better explain what are the exchanged data between each modules within the AquiFR platform.

Comment: *5. It seems that all the models have been recalibrated in order to be able to use the recharge coming from SURFEX. However P13L9-11 confuses me on this point. Have the models been recalibrated in order to use SURFEX as forcing?*

Response: Yes, some information on the need of such calibration was stated page 4 lines 14-17, and is now even made clearer: "*the combined use of SURFEX and SAFRAN provides a consistent set of hydro-meteorological data over an 8 km resolution grid over France, including groundwater recharge and surface runoff from SURFEX, as well as potential evapotranspiration, precipitation, and temperature from SAFRAN. The use of these SURFEX 8 km resolution fluxes made necessary the recalibration of the hydrogeological models included in the platform*". Indeed, it was found that most often, there are some differences between the fluxes estimated by SURFEX and by the original water balance scheme using P/PET, mostly in terms of dynamic. Such differences affected some comparisons between observed and simulated heads, either positively or negatively. To give more information on this recalibration, a new subsection is now added in the new "Methodology" section 3 presented at the end of this document.

Comment: *6. It is not discussed if the recalibrated models forced by SURFEX perform better or worse than the same models, calibrated with P/ETP data and using P/ETP data as forcing. What is the impact of using SURFEX as forcing? Having a homogeneous forcing has value, but does it have downsides?*

Response: This is a good question, and it is true that no information was given in the first version of the article. Overall, the statistical results obtained with SURFEX were similar to those obtain with the original version. A sentence is now added to stress out this point (see answer above), and some information is added in Table 1. Such result is indeed disappointing, as SURFEX is a more physical model and is more demanding computationally. It is one objective of the AquiFR project to improve such results.

Comment: *7. How good is the partition of surface runoff and drainage of SURFEX, in general? This is a key input for the whole system, but it is not validated, not even discussed in the paper. As far as I understand SURFEX may have some empirical parameters in order to determine surface runoff. Has*

this been calibrated? I would like to see a discussion (and data if possible) on the quality of the SURFEX recharge, as it is the main input for the hydrogeological models used in AquiFR.

Response: The SURFEX partition of the surface runoff and drainage may differ from those calculated by the original models. However, it is difficult to distinguish which of the two is closest to the truth, since the truth is unknown, and as, after recalibration, the statistical results obtained by the two versions are similar. It was necessary to modify the partition between surface runoff and drainage only for the Somme basin by using the total runoff. Comments on this point are now added in section 3 (provided at the end of the text) as well as in Table 1. Detailed information is provided in a report accessible online (Habets et al., 2017).

Comment: *8. In the Somme river you don't use SURFEX's partition between runoff and recharge. It seems, that GARDENIA (no citation is provided) adds them together and makes a new partition. Why? How? This should be explained.*

Response: It is now clearly stated that the partition between surface runoff and groundwater recharge in the Somme basin was biased by SURFEX with an overestimation of surface runoff in the North and an underestimation in the South. GARDENIA is the name of the water balance scheme used originally in MARTHE. But, to avoid confusion, we removed the name, added a reference, and some explanations on how it works: " *In order to compensate for this imbalance the total runoff provided by SURFEX was split into surface runoff and groundwater recharge using the original water balance scheme of MARTHE. This water balance scheme is based on a reservoir approach (Thiéry, 2014), for which the parameters were calibrated. Only one reservoir was used, enabling to modify the partition of the surface runoff, and to account for a delay on the groundwater recharge in order to mimic the impact of the deep unsaturated zone*." In details, the reservoir we used is depicted below. H is the head in the reservoir, and is filled by the total runoff from SURFEX. THG is a time transfer coefficient and RUIPER is a partition coefficient that was calibrated. Using such reservoir, not only the partition of the flow between surface runoff and groundwater recharge is modified, but also the dynamic of the flow. This is important in the Somme basin since there is a deep unsaturated zone that is not simulated explicitly in the MARTHE model (see for instance Habets et al., 2010, Multi-model comparison of a major flood in the groundwater-fed basin of the Somme River (France), HESS)

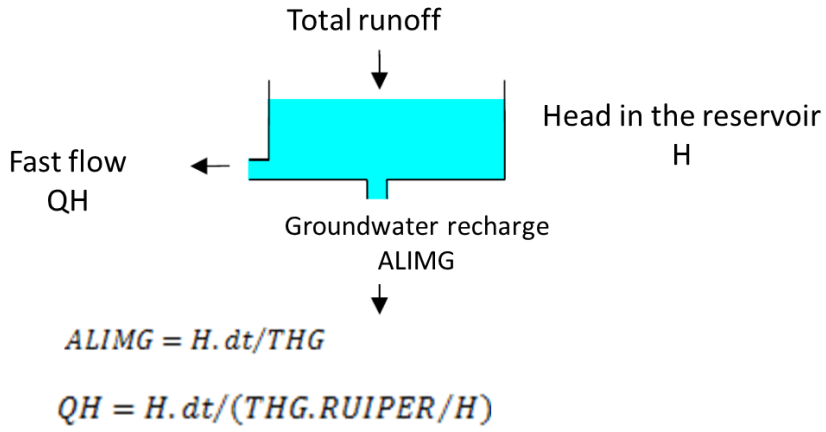

Figure 1 Partition of the total runoff of Surfex in the MARTHE Somme basin

The figure below presents the comparison of the river flow observed and simulated with Surfex with and without a new partition of the total runoff.

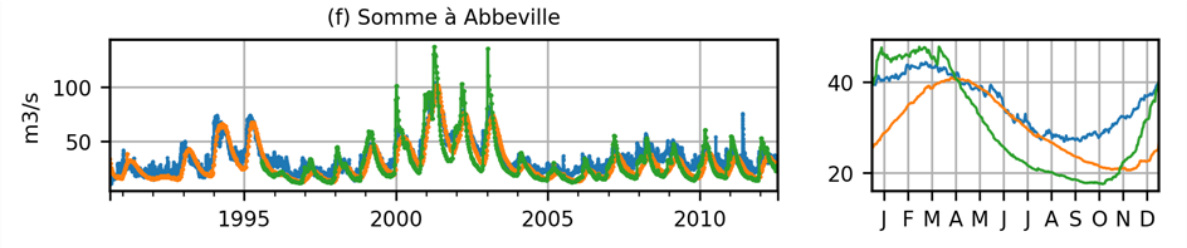

Figure 2 Comparison of the river flows at the outlet of the Somme basin between observations (blue) raw simulation with SAFRAN-SURFEX (orange) and simulation with the total runoff estimated by SAFRAN SURFEX and a partition of the surface runoff and drainage based on the MARTHE original water balance scheme

Comment: *9. Some applications of MARTHE need observed streamflow as an input (boundary conditions). How will you simulate climate change in this area? Why don't you use model streamflow? You simulate it, don't you? You should clarify this point.*

Response: In MARTHE, model streamflow are not simulated outside the simulated aquifer domain. Therefore, if the model does not encompass the entire river basin, boundary conditions are needed to impose flow on these rivers. We used observed streamflow in this version of AquiFR, but it is planned to use a new modelling method based on a lumped-parameter rainfall-runoff model to provide upstream river flows. The text is now modified: *"In the near future, the advantage to have the atmospheric forcing and surface fluxes over the entire domain will be used to estimate the upstream flow based either on a lumped-parameter rainfall-runoff model integrated in the MARTHE computer code or by the RAPID river routing model using a fine scale river network covering all France."*

Indeed, we have a hydrographic network over France, that is used for instance in SIM, but it has a 1km resolution, which is often not enough to match with rivers that are not fully simulated in the hydrogeological models.

In the climate change simulation we have done yet, the hypothesis is to have stable boundary conditions. Therefore, the flow of these not-fully simulated rivers, but also, the sea level, and the surface and groundwater abstractions are expected to be the same as in present day. Of course, it is clear that these hypotheses are not valid, and that the results only provide a first order impact of climate change.

I also have some questions on the cal/val procedure.

Comment: *1. Have you calibrated all the models over the same time period? If no, why? Due to data availability?*

Response: That is correct. The choice was made to calibrate on the same period used by the original model. This ensures that all the data needed are available, and allows comparing fairly with the original models. Please, report to the new section 3.2 provided at the end of the document.

Comment: *2. Do you validate all the models over the whole 60 year period? Do you use the calibration data also for validation? Do you only validate on independent data? The text is not clear to me on this regard and this is a very important issue. Not only for heads, also for streamflow. A model should not be validated using the same data it was used for calibration. If this cannot be avoided, it must be well justified.*

Response: The models are evaluated over the whole 60 year period. As described in the new Methodology section, the calibration procedure was done for each model using the same calibration

period that were used to develop each model (see references in Table 1 and (Habets et al., 2017)). The new methodology section helps to better explain this. However, the validation presented in the article covers the 60-year period, restricted to the availability of the observation. Thus, the calibration and validation periods are different, but the validation period encompasses the calibration period. As all the models were not calibrated on the same period, and as the temporal availability of each measurement varies, it was the only way to have a full assessment of the whole AquiFR platform.

Comment: *3. You show the metrics you used for validation, but not for calibration. I guess that each model is calibrated differently, using different tools. Is this the case? This should be commented.*

Response: All the models were calibrated using the same statistical criteria: Efficiency, correlation and ratio for stream flow, and RMSE and biases for piezometric heads. As stated in the article, no automatic calibration tools were used, but only the skill of hydrogeological experts. These two points are now more clearly stated in the new 3. Methodology section : *"Hydrodynamic parameters, including hydraulic conductivities and specific yields, were modified based on hydrogeological expertise in order to obtain the best fit between observations and simulations. The calibration was made only on the piezometric heads, except for the MARTHE Somme model for which piezometric heads and riverflows were accounted for, and for the kartsic systems with karst spring flows only. All the models were recalibrated using the same statitiscal criterias."*

Comment: *4. You also validate using the NSE. Have you considered the KGE? Or even better, the non parametric version of the KGE (Pool et al, 2018)? The KGE allows to separate the contribution of the correlation, the bias and the standard deviation. The non parametric form makes less assumptions on the underlying data distribution so it can be used with different kinds of variables with less problems. Also, the non parametric form is less sensitive to extremes (so you would not need to calculated the sqrt of the streamflow, as you do). I guess it is too late to change this, but you should consider this in the future.*

Response: Thank you for this comment. Indeed, it is true that the KGE has some advantages compared to the NSE. This is a point that we will consider in the future. A sentence is added in the discussion : "*Some statistical scores using less assumptions on the underlying data distribution, such as the non-parametric variant of the Klunge-Gupta efficiency score, could be used to reduce the sensitivity to the extremes (Pool et al., 2018)."*

Comment: *5. Could you explain with more detail what is the NRMSE-BE? Have you substracted the mean and divided by the standard deviation and then calculated the RMSE? Have you removed the seasonal signal? A little bit more detail on this unusual metric should be provided*

Response: The details are now given in the new subsection 3.2 (see the methodology section at the end of this document).

Comment: *1. Which method do you use to calculate the standardized series? Is it parametric or non parametric?*

Response: The calculation of the Standardized Piezometric Level Index is similar to the calculation of the Standardized Precipitation Index (SPI) (McKee et al., 1993). The SPLI is an indicator used in the Monthly Hydrological Survey published each month. Details about its computation are given in Seguin, (2015). Considering a piezometric head time series of N years, the steps are the following:

- Step 1 : the monthly mean observed time series is computed

- Step 2 : constitution of twelve monthly time series (January to December) over the N year period. For each time series of N values, a non-parametric kernel density estimation (KDE) allows estimating the best probability density function (pdf) fitting the observed histograms. The SPI uses a gamma distribution, but time series of piezometric heads show a big variety of histogram. Therefore, the use of a KDE to estimate a pdf fitting the observed histogram is preferred.
- Step 3 : For each month from January to December, the adjusted pdf is projected over the standardized normal distribution using a quantile-quantile projection.

Figure 3 of the present document shows the procedure for the Omiécourt piezometer. The KDE helps to obtain a fit of the probability density function from the observed histogram. The cumulative density function is deduced, and a projection over the standardized normal distribution allows deducing the SPLI.

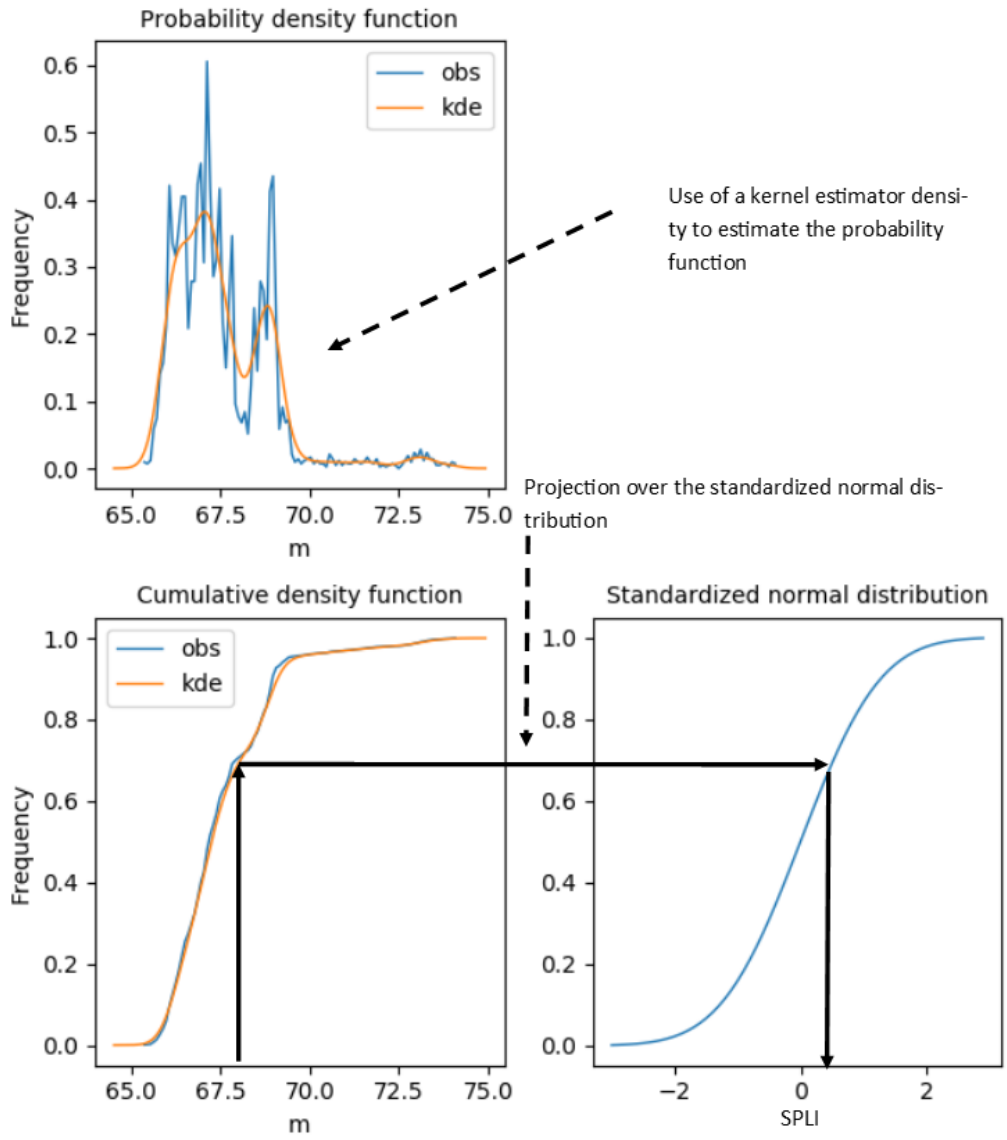

Figure 3: Computation of the SPLI for the Omiécourt piezometer. The probability density function is estimated using kernel density estimator from observed monthly piezometric head values. The estimated cumulative density function is then estimated from the fitted pdf, and a quantile-quantile projection on the standardized normal distribution allows computing the SPLI.

The authors agree that the presentation of the SPLI was not detailed enough. A new presentation is proposed in the new section 3.3 Methodology.

Comment: *2. If it is parametric, which distribution do you fit your data to? Does it fit to all areas equally well?*

Response: As the observed distribution depends on the area where the piezometer is located, a non-parametric KDE is used to estimate the best pdf to fit the observations (Seguin, 2015).

Comment: *3. Figure 10 shows the distribution of the different categories of the SPLI. But some of them are bimodal. I would expect a normal distribution as a standardized variable involves renormalizing the data to a normal distribution. Why these figures don't show a normal distribution?*

Response: Figure 3 of the present document shows the histogram of the observed values which is bimodal. The fitted pdf from KDE is also bimodal, and therefore its projection on the standardized normal distribution will keep this bimodal characteristic.

Comment: *I suggest adding a Methodology section where the cal/val procedure is presented and where the indicators (NRMSE-BE, NSE) and standardizations (SPLI) are presented.*

Response: A new methodology section is added to the revised manuscript, including a presentation of these indicators.

Comment: *Anthropic processes: You take pumping into account for some models. But the subsequent irrigation is not taken into account by SURFEX. Can you comment a little bit more on the current state of anthropic impacts in AquiFR and how this affects the results?*

Response: Most of the groundwater abstraction is used for drinking water. Crop irrigation is not taken into account in the present version of AquiFR. This process can be activated in SURFEX. However, it involves setting up strong hypotheses (where are the irrigated fields, what are the irrigated volume for each field, and when is the irrigation provided) that are beyond the scope of the purpose of the evaluation proposed in this paper. As for the bidirectional coupling between groundwater and SURFEX, this is an option that could be used in the future development of AquiFR.

Specific comments

* P2L6: "Thus, modeling is still a useful tool ...". Well, even with high resolution remote sensing data of storage in aquifers, models would still be useful, as they allow to connect aquifers with the rest of the system (soil, streams, etc.).

Response: The author agree. This sentence is changed into *"At these regional scales, modeling can be a useful tool to provide meaningful information on the groundwater resources (Aeschbach-Hertig and Gleeson, 2012)."*

* PL1: "3 groundwater flow software" -> 3 groundwater flow models.

Response: We try to distinguish software (numerical code) from models (regional models). For example, MARTHE is a hydrogeological modeling software, and 5 models have been developed using this software: the Somme, Poitou-Charentes, Nord-Pas-de-Calais, Basse-Normandie and Alsace models.

"3 groundwater flow software" is replaced by "3 hydrogeological modelling software"

* P4L20: "period.In" -> period. In

* P6L17: "gathersnumerical" should be separated.

* P6L29: coupledto should be separated.

Response: All these remarks are now corrected.

* P7L18: "set of rivers organized in sub-basins". Is this the basis of the acronym? I guess it is in its French form. Maybe it would be better to just put the French name.

Response: It is the english translation of the French acronym. We kept the french name, and as a consequence, we do the same for all the other acronyms that is SAM and MARTHE.

* P8L14: GARDENIA (citation needed). You should also explain how GARDENIA works.

Response: We decided not to provide the name GARDENIA and only to keep the reference to the simplified water balance scheme that is implemented in MARTHE. It is effectively true that this water balance scheme is the same in the GARDENIA software, but it is fully implemented in the MARTHE software and it is now part of MARTHE. The new sentence is: *"This water balance scheme is based on a reservoir for which parameters are calibrated in order to compute the main components of the surface water budget (Thiéry, 2014)."*

* P9L17: observationsat should be divided.

* P11L8: It sensitivity -> Its sensitivity.

Response: All these remarks are corrected.

* P12L18-20: So you validate on the same stations you used for calibration. Do you?

Response: Yes we do.

* P12L33-P13L1. You calculate the sqrt to avoid an excecive influence of extremes. Is this the case? You should explain it.

Response: Yes. It is explained P12L17 to P12L20. We add a brief reminder about this.

* P13L10-11: Here you imply that you didn't recalibrate the models in order to use SURFEX as forcing. But earlier it seems you did. Did you?

Response: Yes we did. See previous answer and section 3.2

* P13L16-29: I would move this into the introduction.

Response: We decided to keep this part in the discussion in order to better highlight the choice of gathering several models in AquiFR as previously shown in section 2.

* P14L6: Which periods were used for calibration?

Response: Periods for calibration were those initially used for calibrated each model independently. This is now better explained in the new Methodology section included in the revised manuscript. This particular part of the discussion

* Fig1: What do the arrows mean? What fluxes are send to the post-processing and what is send back to SAFRAN/SURFEX? I would add labels to the arrows.

Response: The arrows illustrated the flux exchanges. The new proposed scheme better explains this.

* Fig7: Put the legend outside of the first plot.

Response: The legend is now outside of the first plot.

* Fig10: Being standardized values, I would expect a normal distribution, but on three cases it is bimodal.

Response: A new explanation of the SPLI indicators in the new methodology section helps to understand this.

* Fig11b: difficult to see the circles.

Response: The background SPLI map is now more transparent in order to better highlight the circles.

* Fig12: Why are the x-axis time scales so different? Is it related to data availability? Which is the calibration period?

Response: x-axis time scales are different because the axis limits are related to the observed data availability which is different for each karstic system. The calibration period corresponds to these axis limits. In order to be consistent with the evaluation of AquiFR, all the 60-year time serie is now shown.

New and Modified figures in the article

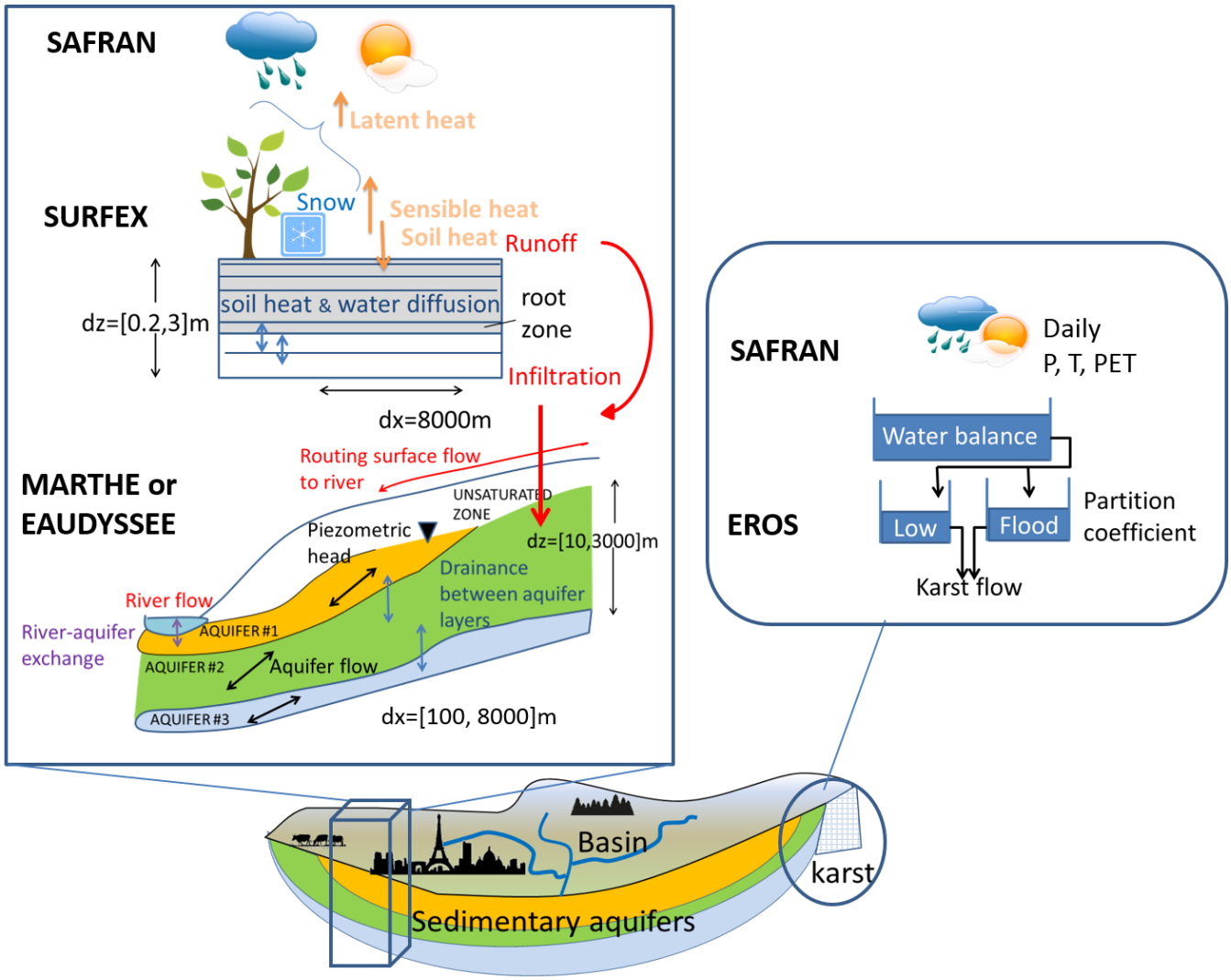

Figure 3: Scheme of the AquiFR physical system. The simulation of the watersheds depends on its hydrogeologic characteristics. For sedimentary basins, the transfer of water within the watersheds is estimated by MARTHE or EauDyssée. It accounts for flows in the unsaturated zones, to (red thin arrow) and in the rivers, in (black arrows) and between (blue arrows) aquifer layers, as well as the exchange between the river and the aquifer (purple arrow). The temporal resolution is daily and the spatial resolution varies from 100 m to a maximum of 8000 m. The depth of the deepest aquifer layer can reach locally about thousand meters. The 8 km spatial partition of the flow between surface runoff and groundwater recharge (red thick arrows) is estimated by the SURFEX land surface scheme. It solves the water and energy budget at a 5 minutes time step. It accounts for the local type of vegetation and soil, the presence of snow, and a multilayer soil that can reach a depth of 3 meters. The atmospheric forcing is provided by SAFRAN. For the karstic systems, the EROS conceptual model is used. It represents each karstic system as lumped basins based on a reservoir approach at a daily time scale. The incoming atmospheric forcing is provided by SAFRAN.

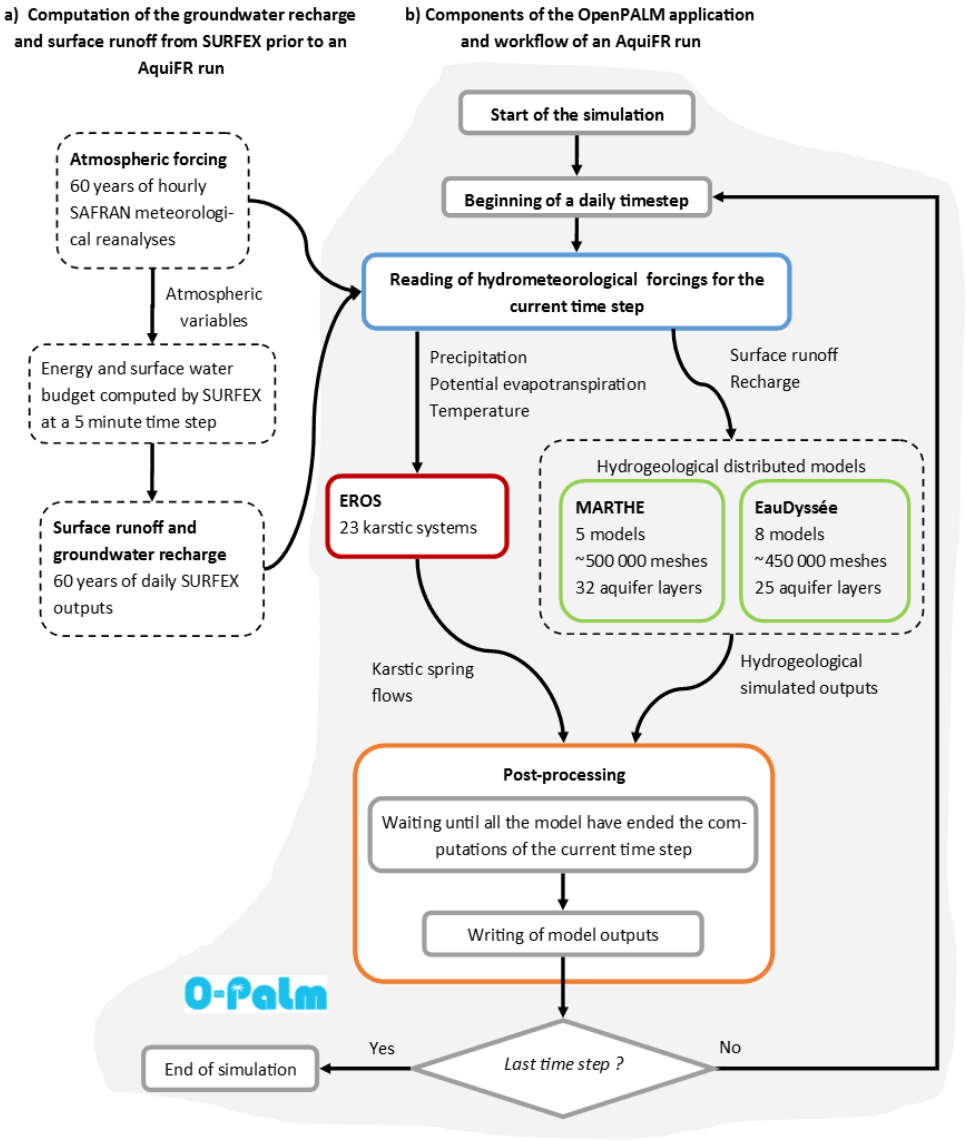

Figure 4: Scheme of the numerical implementation of AquiFR. (a) SAFRAN and SURFEX are run separately, as well as the processes that extract the daily surface runoff and groundwater recharge at 8 km resolution on a daily time step over the full 60 year period. (b) The components implemented within the coupling system O-Palm are presented. Pre-processing in blue gives access to the surface runoff and groundwater recharge as well as atmospheric forcing to the 3 groundwater models for the current time steps. Then, each hydrogeologic software runs all of their models for the current time step. The fluxes and state variables are then transferred daily to the post-processing, that write the model outputs and manage the following time step.

3 Methodology

3.1 The regional models implemented in the AquiFR platform

AquiFR aims at covering all groundwater resources in France. Figure 2 shows the main aquifers covering France classified by geological type as defined in the French hydrogeological reference system BDLISA (https://bdlisa.eaufrance.fr/). The current version of AquiFR gathers 13 spatially distributed models corresponding to regional single or multilayer aquifers (Table 1 and Figure 3).

Some regions are simulated by two spatialized models (Figure 3): the Somme and the Basse-Normandie basins are covered by MARTHE and EauDyssée models, and the chalk aquifer of the Seine basin is covered by both the EauDyssée Seine model and four EauDyssée sub-models (Marne-Loing, Marne-Oise, Seine-Eure, and Seine-Oise regional models, see Figure 4). This allows a multi-model approach, which can be useful for forecast and climate change impact studies. For these regions, the results presented in this paper correspond to the models that were considered as the best calibrated with the SURFEX fluxes. It corresponds to the four EauDyssée sub-models over the Seine basin and the Somme and Basse Normandie MARTHE models. Figure 3 also shows the 23 karstic systems (median catchment area of 99 km2) simulated by EROS (Thiéry, 2018b) as well as the hard rock aquifer in Britany that will be simulated using a hillslope model (Courtois, 2018; Marçais et al., 2017) and integrated in the near future.

Groundwater withdrawals are integrated as input data in the spatially distributed models. On annual average and with respect to the total surface area of the simulated domain, it corresponds to about 16 mm/year (2.4 billion of cubic meters per year) distributed in more than 16 000 grid cells. Data on groundwater pumping are provided by the regional water agencies on the basis of tax reporting. Pumping concerns drinking water, agriculture, and industrial use. The quality of the data set as well as its temporal extension varied for each regional modelling, although the latter does not exceed 20 years. Further details on regional models can be found in the references listed in Table 1. To extend the pumping estimation to the 1958-2018 period, a monthly mean annual cycle is used for the years without data. River routing is performed based on kinematic wave approach in MARTHE and by the RAPID model based on the Muskingum approach (David et al., 2011) in EauDyssée. River-groundwater exchanges are in both directions for all the models. Each regional model uses its own river network at its own resolution. Most of the simulated domains encompass the entire river basins corresponding to the simulated rivers. Only the Alsace and the Poitou-Charentes basins are partially represented. Therefore, they need to prescribe time dependent boundary conditions at the upstream of some rivers based on river flow observations. If the observed data don't cover the full period, the missing values are filled by the daily mean annual observed river flow. In the near future, the advantage to have the atmospheric forcing and surface fluxes over the entire domain will be used to estimate the upstream flow based either on a lumped-parameter rainfall-runoff model integrated in the MARTHE computer code or by the RAPID model using a fine scale river network covering all France.

3.2 Calibration of the hydrogeological models

The original hydrogeological regional models were developed independently most often based on stakeholder requests. The water budgets in these models were usually computed using less physical methods and atmospheric local data (precipitation and temperature) that differ from the physically-based approach using SURFEX and SAFRAN. As a result, in order to be consistent with the estimation of the groundwater recharge estimated by SURFEX, most of the regional models were recalibrated based on the SURFEX fluxes (Habets et al., 2017). This recalibration effort was not undertaken for the Alsace and Loire models since both of them will be soon updated and then recalibrated.

Periods of recalibration were the same as those initially used to develop and calibrate each model (see references in Table 1), in order to facilitate the comparison between the recalibrated models and the initial models. Hydrodynamic parameters, including hydraulic conductivities and specific yields, were modified based on hydrogeological expertise in order to obtain the best fit between observations and simulations. The calibration was made only on the piezometric heads, except for the MARTHE Somme model for which piezometric heads and riverflows were accounted for, and for the karstic systems with karst spring flows only. All the models were recalibrated using the same statistical criterias. A comparison between the initial water budget of the models and the SURFEX fluxes was performed as a first step to estimate the need for recalibration of each model.

Some models, such as the Seine EauDyssée model, were not recalibrated since they perform equally well with the use of the SURFEX fluxes (see Table 1). In contrast, the MARTHE Somme river basin model was characterized by an excess of surface runoff in the north and a deficit to the south. In order to compensate for this imbalance, the total runoff provided by SURFEX was split into surface runoff and groundwater recharge using the original water balance scheme of MARTHE. This water balance scheme is based on a reservoir for which parameters are calibrated in order to compute the main components of the surface water budget (Thiéry, 2014). Only one reservoir was used, enabling to modify the partition of the total runoff and to account for a delay on the groundwater recharge in order to mimic the impact of the deep unsaturated zone. This improved the simulation of the river flows using the SURFEX total runoff. Once the new partition was estimated, the permeability was recalibrated. The Somme basin is the only one for which only the total runoff from SURFEX was used. For the other basins, the estimation by SURFEX of the partition of the water fluxes between surface runoff and groundwater recharge was used. Overall, the performance of models using the fluxes from SURFEX are similar to the original version, although locally, they may be better or otherwise degraded.

For the karst system software EROS, the models were calibrated based on the SAFRAN atmospheric analysis by using an optimization of the statistical comparison between observed and simulated daily river flows.

More information about the method of recalibration is given in Habets et al. (2017).

3.3 Evaluation criteria of the 60 years long-term simulation

Statistical criteria are used to evaluate the long-term simulation. The bias allows evaluating the relative mean deviation between the observation and the simulation. It is calculated as follows:

$$BIAS = \frac{1}{n}\sum_{i=1}^{n}\left(X_{obs}(t) - X_{sim}(t)\right), \quad (1)$$

with n the number of observed values, $X_{obs}(t)$ and $X_{sim}(t)$ the observed and simulated values respectively at time t.

The Root Mean Square Error (RMSE) score allows estimating the differences between the observed and simulated values. It is often used to compare observed and simulated piezometric heads. However, the computation of the RMSE score is strongly affected by the biases. Therefore, we computed a RMSE bias-excluded score in order to better assess the simulation in terms of amplitude and synchronization. Moreover, this RMSE bias-excluded score is normed with respect to the observed standard deviation for each observation. It takes into account the differences of variability between the observed points and to better compare them with each other. This normed RMSE bias-excluded (NRMSE_BE) is expressed as follow:

$$NRMSE\_BE = \frac{1}{\sigma_{obs}} \sqrt{\frac{\sum_{i=1}^{n}[(X_{sim}(t)-\overline{X_{sim}})-(X_{obs}(t)-\overline{X_{obs}})]^2}{n}} \quad (2)$$

with $\overline{X_{sim}}$ the temporal mean of simulated values over the considered period and σ_{obs} the observed standard deviation.

The Nash-Sutcliffe model Efficiency score NSE (Nash and Sutcliffe, 1970) measures the variance between the observed and simulated values. It is often applied to compare observed and simulated river flows but can be used for other variables. Its sensitivity to high-frequency fluctuations makes its use for comparing groundwater levels less obvious. This criteria is equal to 1 when the model fits perfectly the observations. A NSE above 0.7 is generally accepted as a good estimate of the signal dynamic, however depending on the hydrogeological and climate context of the basin. A negative NSE means that the mean observed signal is a better predictor than the model. The NSE is calculated as follows:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(X_{obs}(t)-X_{sim}(t))^2}{\sum_{i=1}^{n}(X_{obs}(t)-\overline{X_{obs}})^2}, \quad (3)$$

with $\overline{X_{obs}}$ the temporal mean of observed values over the considered period.

One way to evaluate the ability of the simulation to capture extreme events is to use the Standardized Piezometric Level Index (SPLI). The SPLI is an indicator used to compare groundwater level time series and to characterize the severity of extreme events such as long dry period or groundwater overflows (Seguin, 2015). Assessing the ability of the AquiFR modelling platform to reproduce this indicator is important since the main objective of this platform is to predict such extreme events in short-to-long terms hydrogeological forecasts for groundwater management. The SPLI indicator is based on the same principles as the Standardised Precipitation Index (SPI) defined by (McKee et al., 1993) to characterize meteorological drought at several time scales. First, monthly mean time series are computed from a time series of piezometric heads. Then, twelve monthly time series (January to December) are constituted over the N years of the time series period. For each time series of N monthly values, non-parametric kernel density estimation allows estimating the best probability density function (pdf) fitting the histogram of monthly values. At last, for each month from January to December, a projection over the standardized normal distribution using a quantile-quantile projection allows deducing the SPLI for each value of the monthly mean time series of piezometric heads.

The SPLI values most often range from -3 (extremely low groundwater levels corresponding to a return period of 740 years) to +3 (extremely high groundwater levels). The SPLI allows representing wetter and drier periods in a similar way all over the French national territory.