

3

Material and Methodology

In this study, various methodologies were employed to address the response of the complex interactions between SW and GW systems to the induced recharge through injection wells. The hydrological modeling component utilizes the Soil and Water Assessment Tool (SWAT) to simulate key hydrological processes, including overland flow, river flow, and GW recharge. SWAT's ability to handle diverse land uses and climatic conditions makes it suitable for understanding watershed dynamics. The model operates through a well-defined input-output (I/O) structure and is calibrated using the Sequential Uncertainty Fitting algorithm (SUFI2) to reduce uncertainty and improve the accuracy of predictions. A three-dimensional GW flow equation is solved using the Finite Difference Method (FDM) with MODFLOW in GMS software for GW modeling. This approach captures the complexity of subsurface flow, accounting for heterogeneity in aquifer properties and interactions with SW. The model's calibration is performed using a combined method using the Parameter ESTimation (PEST) tool for aquifer parameters and Particle Swarm Optimization (PSO) for stream bed conductivity, which helps fine-tune the parameters to achieve a better fit between observed and simulated data, thereby enhancing the model's reliability.

To complement these models, machine learning techniques, including Random Forest Regressor (RFR), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM) networks, are integrated based on the application and relative accuracy achieved. These Machine learning (ML) models have been used for performing downscaling of climate variables (Temperature and Precipitation for four IPCC CMIP6 Scenarios i.e.,

SSP126, SSP245, SSP370, and SSP585) and GW storage anomalies (GRACE-GWSA) during the assessment of GW availability and climate change impact assessment in the Varuna River Basin (VRB). These ML models have also been used to predict hydrogeological variables (such as stage-discharge relationship, inlet discharges in the SWAT model, and GW heads at boundary cells) to build the forecasting MODFLOW model.

The outputs have been analyzed using various tools and techniques. In analyzing temporal patterns, the study employs trend and change-point analysis methods to detect shifts and trends in hydrological and GW data. The Mann-Kendall (MK) test and Sen's slope estimator identify statistically significant trends, while change point analysis highlights abrupt shifts. Sensitivity analysis is conducted to evaluate the influence of input parameters on the developed novel methodologies. Both local and global sensitivity analyses are performed, with Sobol's method being used to assess each parameter's contribution to the output variance. Finally, the accuracy of the models is evaluated using multiple metrics, including the coefficient of determination (R^2), root mean square error (RMSE), and Nash-Sutcliffe Efficiency (NSE). The common methodologies employed in the research work are explained in detail in this section, while the novel methods are explained in the corresponding section.

3.1 HYDROLOGICAL MODELLING WITH SWAT

The hydrological drivers have been simulated with the SWAT model (Arnold, J. et al., 2012). The SWAT is a physically based, distributed hydrological model used to analyze watersheds at the basin scale. It uses physical relationships to simulate various hydrological processes within a river basin, such as rainfall-runoff, sediment transport,

and nutrient cycle. The regional-scale watersheds (or basins) are divided into sub-watersheds/subbasins to capture the heterogeneous soil properties and land uses. The input information for each subbasin is divided into five categories: climate, hydrologic response units (HRUs), ponds/wetlands, GW, and the major channel that drains the subbasin (reach). HRUs are sub-basin land sections with distinct land cover, soil, and management combinations (Arnold, J. et al., 2012).

The water balance is the main driving force in SWAT for simulating every process (sediment and nutrient transport). The hydrological simulation of the watershed has been carried out in two phases. The first was the land phase of the hydrologic cycle, and the second was the routing phase, which defines the movement of water, sediments, and nutrients in the stream network.

3.1.1 Overland flow phase of the SWAT model

The overland flow phase of the hydrologic cycle is based on the water balance equation:

$$SW_t = SW_0 + \sum_{i=1}^t (R_{day} - Q_{surf} - E_a - w_{seep} - Q_{gw}) \quad 3.1$$

Where SW_t is the final soil water content (mm), SW_0 is the initial soil water content (mm), t is the time (days), R_{day} is the amount of precipitation on the day i (mm), Q_{surf} is the amount of surface runoff on the day i (mm), E_a is the amount of evapotranspiration on the day i (mm), w_{seep} is the amount of percolation and bypass flow exiting the soil profile bottom, on the day i (mm), and Q_{gw} is the amount of return flow on the day i (mm). All the units are represented in the equivalent water depth in mm.

SWAT modelling simulates two aquifers within each subbasin: an unconfined, shallow aquifer that contributes to the subbasin's main channel flow and a confined deep aquifer

which is assumed to affect streamflow beyond the watershed boundaries (Neitsch et al., 2011). An aquifer percolation coefficient routes a fraction of the shallow recharge to the deep aquifer in each time step. A revap coefficient partitions the water in shallow aquifers to move in the plant root zone in response to GW deficiency. The overall budget of the GW module of SWAT is given as (Neitsch et al., 2011).

$$aq_{sh,i} = aq_{sh,i-1} + (1 - \beta_{deep})w_{rchg} - Q_{gw} - w_{revap} - w_{pump} \quad 3.2$$

Where $aq_{sh,i}$ is the amount of water stored in the shallow aquifer on day i (mm), $aq_{sh,i-1}$ is the amount of water stored in the shallow aquifer on day $i-1$ (mm), β_{deep} is the deep aquifer percolation coefficient, w_{rchg} is the amount of recharge on day i (mm), Q_{gw} is the base flow on day i (mm), w_{revap} is the amount of water moving into the soil zone in response to water deficiencies on day i (mm), and w_{pump} is the amount of water removed from the shallow aquifer by pumping for irrigation on day i (mm).

3.1.2 Water routing phase of the SWAT Model

In the SWAT model, water routing through the river system is based on the kinematic wave equation. The water flow from one reach to another is calculated using the Muskingum method, a commonly used method in hydrological models to simulate flood wave movement in rivers.

The general equation for the Muskingum method is given as:

$$Q_{i+1} = C_1 \cdot I_i + C_2 \cdot I_{i+1} + C_3 \cdot Q_i \quad 3.3$$

Where Q_{i+1} = outflow at time step $i+1$, Q_i = outflow at time step i , I_i = inflow at time step i , I_{i+1} = inflow at time step $i+1$ and $C1$, $C2$, $C3$ = routing coefficients that are calculated based on channel storage and travel time.

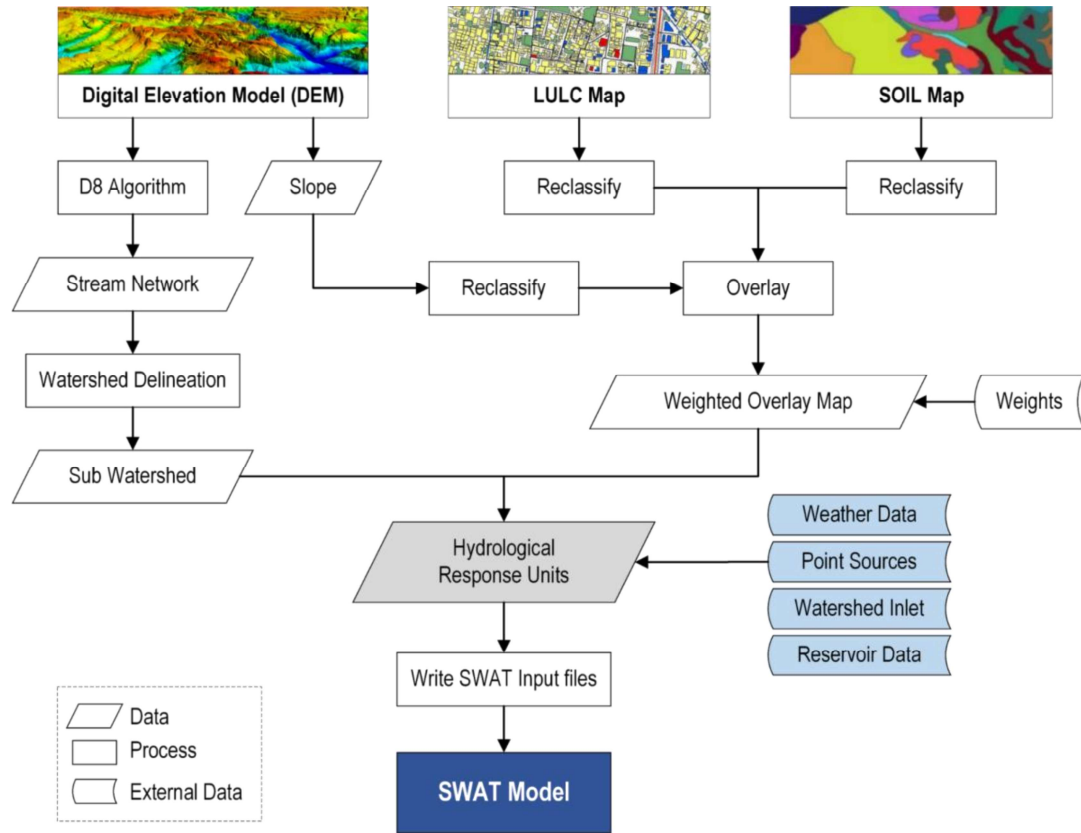


Figure 3.1. Schematic flow chart for hydrological modeling with SWAT

3.1.3 User Interface and Input/Output

One of the major benefits of SWAT is its straightforward graphical user interface integrated with GIS software (QGIS or ArcGIS) using a plugin (QSWAT or ArcSWAT). The integration with the GIS environment makes it easier to use standard spatial data format for land use and soil during conceptualization. The weather generator utilizes the standard ASCII text files as daily weather parameter inputs. A Python tool (IMD_SWAT³)

³ https://github.com/Ranveer82/IMD_SWAT

has been developed to use IMD 2D gridded data to generate input files for the weather generator. The center of the grid has been taken as station locations, and the station elevation has been taken from the digital elevation model (DEM). The point sources and flow from the draining watershed must be defined in the subbasins to which the sources are draining water. The daily/monthly/yearly flux of water sediments and nutrients needs to be in the standard ASCII input formats (Arnold, J., Kiniry, R., Williams, E., Haney, S., Neitsch, 2012) and can be easily prepared using MS Excel. The SWAT model outputs are also ASCII-based and have a tab-separated table structure, which can easily be interpreted using MS Excel (for a smaller number of subbasins) or with the standard library in Python. The step-by-step procedure to build a SWAT model from scratch is illustrated in **Figure 3.1**.

3.1.4 Calibration and Validation

The uncertainties resulting from simplified models, processes not included in the model, and processes the modeler is unaware of make the hydrological modeling process difficult (Abbaspour et al., 2007). The modeling of the given study area has uncertainties due to (1) the accuracy of input data, (2) GW pumping for irrigation and domestic use, (3) biases in observed data, and (4) unknown processes such as river water abstraction. The calibration using a single variable can introduce bias in other hydrological fluxes (Abbaspour, 2019). The multi-site variables (River flow, and evapotranspiration) were generally required to calibrate the SWAT models better.

The Sequential Uncertainty Fitting (SUFI-2) algorithm in SWAT-CUP (Abbaspour, 2019) has been used to calibrate the model with Nash-Sutcliffe efficiency (NSE) as the objective function (**Figure 3.2**). In SUFI-2, parameter uncertainty is represented as ranges (uniform distributions) and includes all sources of uncertainty, such as those linked to driving variables (e.g., precipitation), the conceptual model, parameters, and measured data. As

these parameter errors propagate through the model, they cause uncertainties in the model's output variables, which are represented as 95% probability distributions. These distributions are based on the 2.5% and 97.5% percentiles of an output variable's cumulative distribution, which was computed by propagating parameter uncertainty using Latin hypercube sampling. This range is known as the 95% prediction uncertainty (95% PPU), and it is used as the model output in a stochastic calibration framework. The two measurements, P-factor and R-factor, examine the goodness of fit and the extent to which the calibrated model accommodates uncertainty. The P-factor represents the percentage of observed data covered by our 95PPU model. The R-factor indicates the thickness of the 95PPU envelope. In theory, the P-factor ranges from 0 to 100%, while the R-factor spans from 0 to infinite. A simulation with a P-factor of 1 and an R-factor of zero accurately represents measured data. The calibration is assessed based on the degree of deviation of these measures from the accurate values.

SWAT calibration using the SUFI-2 algorithm is an iterative process aimed at adjusting model parameters to optimize the fit between simulated and observed data. In this approach, ranges are defined for uncertain parameters, which are then sampled using Latin hypercube sampling. Multiple model simulations are run with these sampled parameter sets, allowing for the evaluation of model performance across a wide range of conditions. The goal is to find parameter sets that reduce the error between simulated outputs, such as streamflow and observed data, improving model accuracy while accounting for the inherent uncertainties in both model structure and input data.

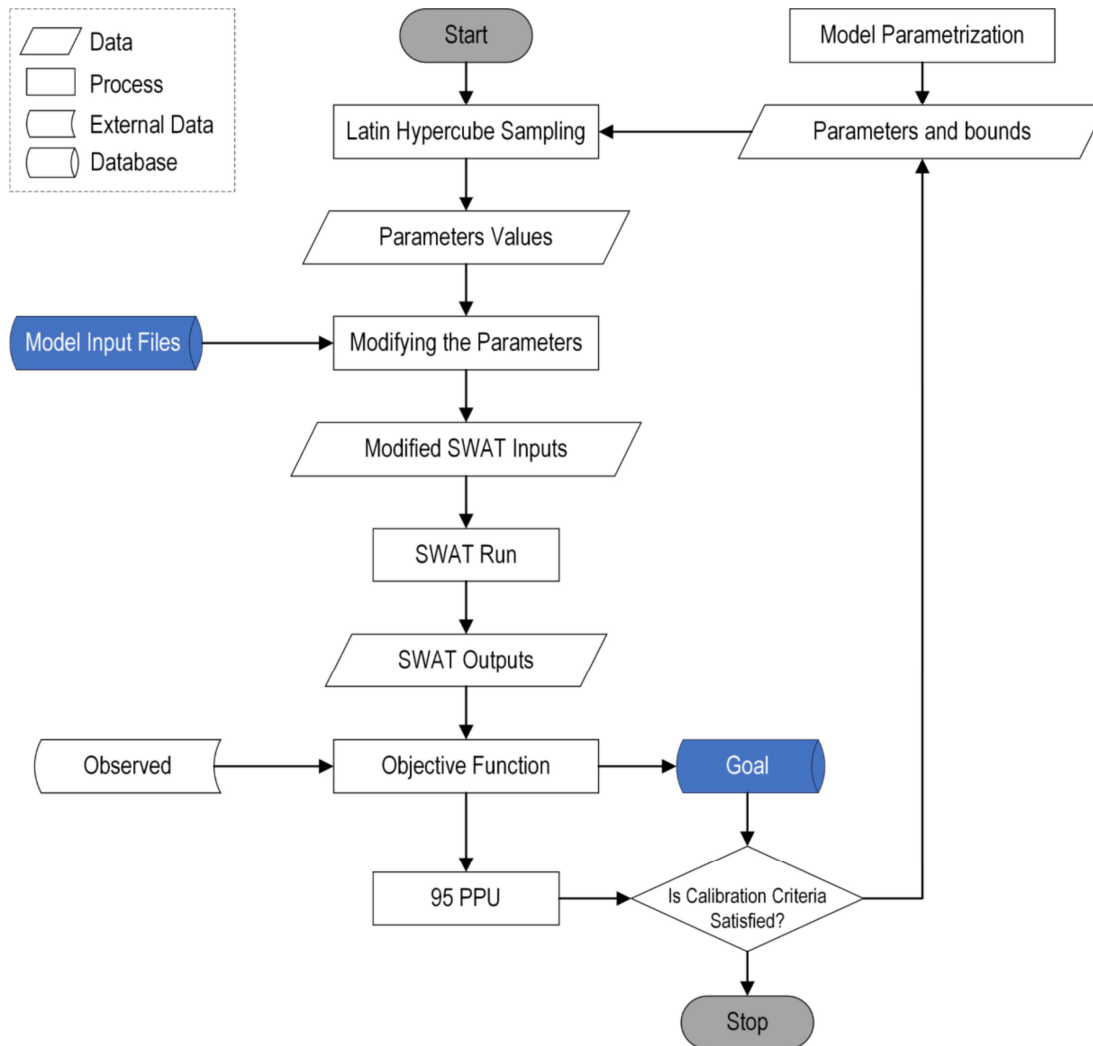


Figure 3.2. Schematic flow chart for Calibration using SUFI-2

3.2 GW MODELING: INTEGRATED METHOD

Modeling GW dynamics in an area is crucial for assessing hydrogeological variables and understanding their response to climatic and anthropogenic disturbances (Anderson et al., 2015). GW models, such as MODFLOW, employ numerical methods like the finite difference method (FDM) to solve GW flow equations, enabling the simulation of complex GW systems (Fetter, 2001). These models are widely used for water resource management, predicting the impact of human activities (e.g., pumping,

irrigation)(Hodges, 2019; Leng et al., 2014; Pokhrel et al., 2015; Wada et al., 2017), assessing GW availability (Menció and Mas-Pla, 2008; P. J. Omar et al., 2020; Schwartz and Ibaraki, 2011), managing contamination (Clement et al., 1998; Mustafa et al., 2016; Panday and Huyakorn, 2008; Prommer et al., 2003; Simon et al., 2015), and evaluating the effects of climate change on GW resources (Ali et al., 2020; Atawneh et al., 2021; Vergnes et al., 2023). By simulating various scenarios, GW models provide critical insights for sustainable management and decision-making in environmental and hydrological studies (Anderson et al., 2015).

3.2.1 GW flow in saturated media

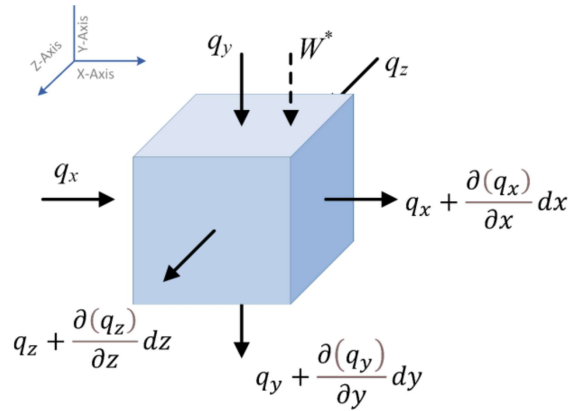


Figure 3.3. Representative elementary control volume of aquifer media

GW flows through porous aquifer media under a variable hydraulic gradient and is generally described by Darcy's Law (Darcy, 1856) for a constant density flow:

$$q = -K\nabla h = - \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \nabla h \quad 3.4$$

Where q is the specific discharge vector (LT^{-1}) and is given as:

$$q = [q_x \quad q_y \quad q_z] \quad 3.5$$

K is the hydraulic conductivity tensor (LT^{-1}); K_{xx} , K_{yy} , and K_{zz} are the hydraulic conductivity along major principal axes x , y and z , respectively; h is the GW head (L); and ∇h is the head gradient vector given as:

$$\nabla h = \left[\frac{\partial h}{\partial x} \quad \frac{\partial h}{\partial y} \quad \frac{\partial h}{\partial z} \right] \quad 3.6$$

Considering an elementary control volume of the aquifer media (**Figure 3.3**), applying water mass balance yields the governing equation for transient flow through a porous heterogenous medium (Anderson et al., 2015):

$$\frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} + \frac{\partial q_z}{\partial z} - W^* = -S_s \frac{\partial h}{\partial t} \quad 3.7$$

Where, q_x , q_y and q_z are specific flux along x , y and z axis (LT^{-1}); S_s is the specific storage (L^{-1}), W^* is the source or sink term depending upon the direction of flow (i.e. flow into or out of the system) (T^{-1}); and t is the time (T).

3.2.2 Numerical Modelling with MODFLOW-NWT

Unlike analytical solutions and the analytic element method (AEM), numerical models use an approximate version of the governing equation (Eq 3.7) to compute hydraulic heads at specified locations. Numerical solutions are not continuous throughout space or time; instead, the head is computed at discrete spatial positions (called nodes) and for specific time intervals (called time steps). However, numerical models can solve the whole transient, three-dimensional, heterogeneous, and anisotropic governing equation (Eq 3.7) with complex boundary and starting conditions (Anderson et al., 2015; Fetter, 2001). This makes the numerical method widely applicable, especially in the case of modelling a complex GW system such as a highly heterogeneous Gangetic alluvial plains

(Das et al., 2021c; P. J. Omar et al., 2020). The detailed concepts and applications of numerical models have been discussed by Anderson et al. (2015).

MODFLOW-NWT (Newton Formulation for MODFLOW)(Niswonger et al., 2011) is a robust version of the widely-used MODFLOW GW modeling software designed to handle complex, nonlinear GW flow problems. It incorporates a Newton-Raphson solver, making it particularly suitable for unconfined aquifers, where drying and rewetting of cells occur due to fluctuating water table levels (Langevin et al., 2021). This feature improves the model's stability and accuracy when simulating dynamic GW systems under varying hydrological conditions. MODFLOW-NWT is often used in applications that require modeling of surface-water/GW interactions (Feinstein et al., 2012; Sophocleous, 2002), highly heterogeneous aquifers (Kang-l, 2014; Ou et al., 2013), and challenging boundary conditions (Niswonger et al., 2011). Allowing the simulation of both steady-state and transient conditions provides valuable insights for water resource management, environmental assessments, and the prediction of GW responses to natural or anthropogenic changes.

Modeling with MODFLOW-NWT requires several steps, which eventually write the input files for the MODFLOW-NWT executable to be read. After defining the modeling purpose, the model conceptualization is performed. This step requires collecting all the important information and datasets to support the modeling objective. The data formats are generally classified as point, linear, or polygon datasets. The classification is related to the extent and uniqueness of data for the selected feature type. The general data used in modeling the GW flow system is summarized in **Figure 3.4**. The next step is to design the GW model, translating the conceptual model into the numerical GW flow model. This step involves designing the grid (spatial discretization), setting boundary conditions, defining aquifer parameters (K, HANI, VANI, Ss, Sy), defining the hydrological stresses

(pumping wells, Evapotranspiration), setting initial condition (starting head), and selecting time steps (temporal discretization). After mapping all the data to the MODFLOW grids, the model is run with an initial set of parameter values. Calibration is the most important step in modeling, and it helps reduce uncertainty related to data measurements and conceptualization (Anderson et al., 2015). The forecasting model requires forecasted data with arguably well-calibrated models (ML or statistical). The prediction for all the stresses and boundary conditions is required and has been performed using different sets of calibrated ML models in this study. It will be explained in the next section. The general modeling steps are illustrated in Figure 3.4, and the corresponding packages to define the stresses and boundary conditions are briefly explained below.

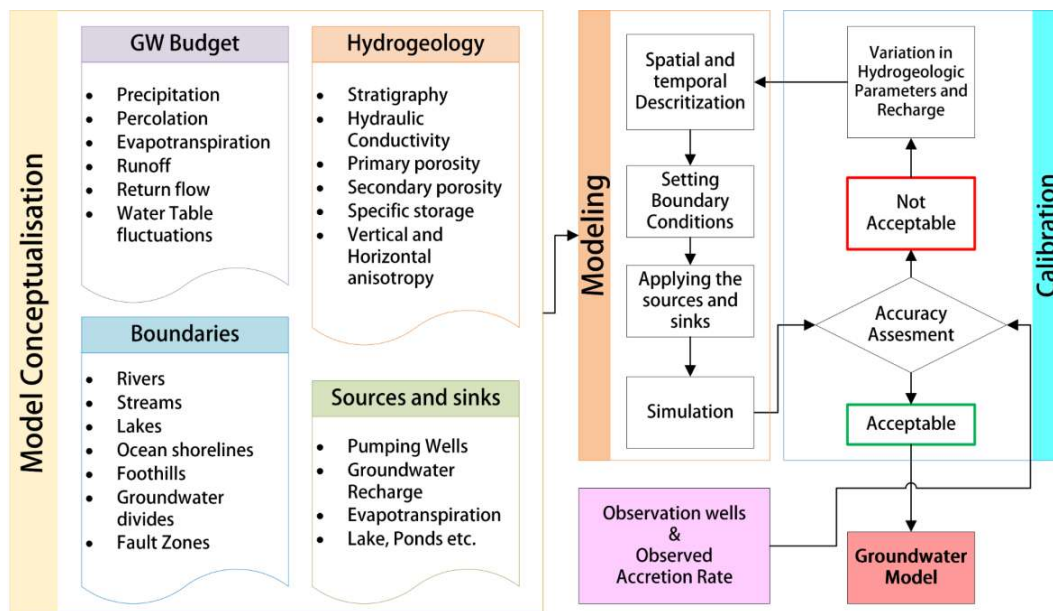


Figure 3.4. GW modeling steps

3.2.2.1 Discretization

Discretization of space and time in the case of GW modeling refers to breaking down the continuous hydrogeological system (aquifer, river, drains, recharge areas, etc.) into grids

of smaller cells and time steps, respectively. This facilitates the approximation of solutions to the complex equations governing GW flow. Discretization is an important step since the number of grid cells directly affects the accuracy of the simulated head (Anderson et al., 2015). The number of cells also affects the runtime and computational requirement as the large number of grid cells will require longer simulation time and large computational resources. The discretization in MODFLOW-NWT is defined by the structural discretization (DIS) package, and the cells are identified by their layer, row, and column position in the grid. The Discretization Packages contain information about cells, such as the cell top and bottom elevations and cell dimensions or areas (in plan view). The Discretization Packages also define cell connectivity. Cell connectivity describes how model cells are connected to one another and the geometric properties of the connections. For the DIS Package, cell connectivity is implicitly described by the grid structure, cell dimensions in the row and column directions, and cell top and cell bottom elevations.

3.2.2.2 Setting Boundaries

The boundary conditions in terms of GW flow models are generally subdivided into two types: 1. Perimeter boundary conditions and 2. Internal boundary conditions (Anderson et al., 2015). Boundary conditions that describe the hydraulic conditions at the edges of the model domain are known as perimeter boundary conditions. Additionally, GW flow models can include boundary conditions representing internal sources and sinks within the system, effectively functioning as internal boundary conditions. The perimeter boundary conditions in MODFLOW-NWT translate the physical GW boundaries, depending upon the flow behavior of GW across these boundaries. Three types of boundary conditions are generally defined;

1. *Specified Head Boundary (Dirichlet Conditions)* where the head along the boundary is set at a known value. The boundary of large SW bodies can act as a specified head boundary (or constant head boundary (CHD) in MODFLOW-NWT) as their water level does not show major changes due to source and sinks for a given time step (Anderson et al., 2015).
2. *Specified flow boundary (SFB) (Neumann conditions)* where the derivative of the head (i.e., q/K by Darcy's law) at the boundary is specified. The no-flow boundary is the special case of this condition, where the specified flow is zero. The bottom boundary of most models is specified as a no-flow condition at contact with the bedrock, which is relatively impermeable. In MODFLOW-NWT, the Well packages (WEL or MNW2) and Recharge package (RCH) are specified flow boundaries.
3. *Head-dependent boundary (HDB) (Cauchy conditions)* allows flow rates to depend on the difference between the head at the boundary and the head inside the model. These boundaries simulate dynamic interactions between the GW system and SW bodies, such as rivers, lakes, or drains. The hydraulic head difference and the conductance of the boundary influence the amount of flow. In MODFLOW-NWT, Stream Packages (STR or SFR2) and River packages (RIV) are specified flow boundaries.

3.2.2.3 Recharge and Evapotranspiration (ET)

Sources and sinks in GW models refer to the flux from the external system, such as recharge, evapotranspiration, or GW extraction through pumping wells. The recharge due to the percolation of rainwater is a primary source of inflow into the aquifer, apart from the boundary flow coming from the upstream areas. The recharge amount primarily depends on the soil characteristics and land use types. However, the general topography

of the area also influences the percolation rate of rainfall (Schiff, 1964). The MODFLOW-NWT recharges on the model's top layer or the topmost active cell, depending upon the modeler.

The Evapotranspiration (ET) Package simulates how water is removed from the GW system through plant transpiration and direct evaporation. It follows these key assumptions: (1) When the water table is at or above a certain level, known as the "ET surface", water is lost from the GW at a constant rate; (2) If the water table drops below a specified depth, called the "extinction depth" or "cutoff depth," evapotranspiration from the GW stops; and (3) Between these two limits, the rate of ET changes linearly based on the water table's depth. The ET surface is defined by the cell's top elevation and the extinction depth based on the plant root zones (Shah et al., 2007).

3.2.2.4 *GW extraction*

GW extraction is a major contributor to the GW budget and needs accurate estimations, especially in areas where the pumping is not monitored (Meza-Gastelum et al., 2022). To tackle data scarcity, the consumptive use method was used in this study, combining crop water and industrial and domestic demands. Agricultural demand was calculated based on the total irrigated area and crop requirements. The domestic demands have been calculated based on the population and per capita demand based on the report of the GW resource estimation committee (Ministry of Water Resource, India). By integrating all demands, the total GW extraction from an area is given as:

$$Q_{i,t} = \left[\sum_j^n CF_j IR_{j,t} \right] \times TCA_i \times \delta_t + GW_{p,i} \times TP_{i,t} + GW_{IN,i} \quad 3.8$$

$$CF_j = \frac{Crop\ Area}{TCA_i} \quad 3.9$$

$$IR_{j,t} = Cu_{j,t} - R_{eff,t} \quad 3.10$$

Where $Q_{i,t}$ is the total water extraction for an area i at time step t ; CF_j is the crop factor of crop type j within the area i ; $IR_{i,t}$ is the irrigation requirement of crop j at time t ; TCA_i is the total cultivated area for i ; δ is the coefficient of irrigational dependency on GW at time t ; $GW_{p,i}$ is the per capita GW demand, $TP_{i,t}$ is the total population for the area i during time step t ; $GW_{IN,i}$ is the industrial demand in area i ; $Cu_{j,t}$ is the consumptive use of the crop j at time t ; and $R_{eff,t}$ is the effective rainfall in the area at time step t .

The GW extraction in MODFLOW-NWT forms a specified flow boundary condition and is managed by the WELL or MNW2 package (Leonard F.Konikow, George Z. Hornberger, Keith J. Halford, 2009). The WELL package simulates the well in the given cell without any criteria of aquifer yield or well losses. The MNW2 package simulates the wells by incorporating the head loss due to well characteristics and aquifer formation. This results in more accurate head estimation in cells with respect to the standard WELL package. As Wells in MODFLOW are simulated as a specified flow boundary, the aquifer only interacts through Wells in terms of flow exchange. For an area with a large number of extraction wells, using MNW2 is not a good strategy since it will create additional complexities, resulting in a larger computational time (Anderson et al., 2015). Therefore, the WELL package was used for GW extraction, and MNW2 was used to simulate injection wells.

3.2.2.5 Rivers and Drains

Rivers act as sources and sinks in the GW flow system, while drains are exclusively modeled as sinks. The interaction of river and aquifer is termed river aquifer exchange (RAE), and this exchange can be spatially and temporally complex (Anderson et al.,

2015). The RAE can be negative or positive based on the nature of the hydraulic gradient between the GW and river stage. The stream reach is called losing if the flow is strictly from the stream to the aquifer and the RAE is positive. The inverse of this phenomenon happens when the stream starts gaining water, and the RAE is negative. The gaining flux from the aquifer to the stream makes the baseflow of the river during the lean period when the contribution of runoff is low.

The rivers are generally modeled based on how they occupy the grid cell. The rivers that fully occupy the cell are simulated as specified head boundaries; however, most of the time, the modeled grid size is larger than the river width and depth. Therefore, the rivers are generally modeled using HDB conditions and conceptualized to be hydraulically connected to the aquifer but do not physically occupy space within the grid. The exchange of water occurs through the stream bed, and the exchanges from the banks are neglected. Several packages are available in MODFLOW-NWT to model rivers (such as River Package (RIV), Stream Package (STR), and Stream Flow Routing Package (SFR2) (Niswonger and Prudic, 2010)). Based on the river characteristics, we have utilized the RIV and SFR2 packages.

- A. **RIV Package:** The River (RIV) Package in MODFLOW simulates the interaction between a river and the GW system. This package models rivers as head-dependent boundaries, where the flow between the river and the aquifer (Q_{GW}) depends on the difference in hydraulic head between the river stage (h_s) and the GW table ($h_{i,j,k}$), as well as the conductance of the riverbed (C).

$$Q_{GW} = -K_{rz} \frac{h_{i,j,k} - h_s}{b_r} LW = C(h_{i,j,k} - h_s) \quad 3.11$$

Where K_{rz} is the hydraulic conductivity of the river bed, b_r is river bed thickness, L is the reach length, and W is the width within the grid cell. The inputs for the

RIV package include the river stages, river bottom elevation, and river bed conductance.

B. **SFR2 package:** The Streamflow-Routing (SFR2) in MODFLOW-NWT simulates the interaction between SW (streams, rivers) and GW, as well as routing streamflow through a network of surface-water channels. The package replaces the older STR package and offers several improvements, including more flexible options for calculating stream depth and streambed conductance. The river cross-section is defined by an eight-point (Figure 3.5) or a wide rectangular channel as required.

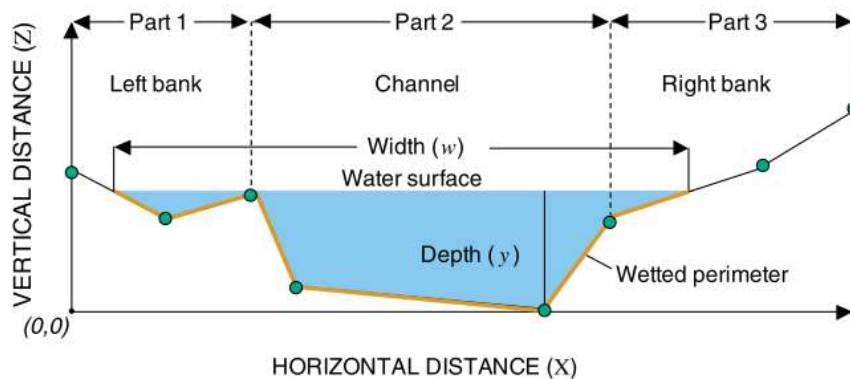


Figure 3.5. An eight-point cross-section is used to compute the depth, width, and wetted perimeter for a stream segment (Prudic et al., 2004)

The SFR2 package solves flow using a continuity equation, assuming steady, uniform flow over discrete time periods. It can simulate flow in streams under varying conditions, allowing for inputs like runoff, precipitation, and evapotranspiration to influence the streamflow within each reach. The package provides multiple methods for calculating stream depth, including Manning's equation and empirical depth-discharge relationships, making it highly adaptable for different stream types. Additionally, it handles solute transport, tracking the

movement of contaminants between the stream and aquifer. The SFR1 package is particularly useful for modeling long-term stream-aquifer interactions and the transport of solutes, making it a valuable tool in managing water resources and studying environmental impacts.

3.2.3 Integration of SWAT output to MODFLOW

The hydrological variables are important for the conceptualization of a 3D-GW Model. Accurate estimation of evapotranspiration and recharge is crucial for reliable model simulation since these comprise a major part of the GW budget. The uncertainty in estimating these variables can highly influence the accuracy of the GW model, which is widely used in the conventional approach (by calibrating the hydrological variable to the observed heads). This uncertainty is minimized by integrating hydrological model outputs into the GW models.

The integration of the calibrated SWAT model into the MODFLOW-NWT packages can be done either by coupling both the models together by a program that links both models (e.g., SWAT-MODFLOW) or by integrating the SWAT model outputs in the MODFLOW-NWT without running SWAT and MODFLOW-NWT together. The latter approach seems appropriate for a regional scale model as the fully coupled models are still under development, and the effort and model complexity required are not fully justified by the accuracy improvement (Haque et al., 2021).

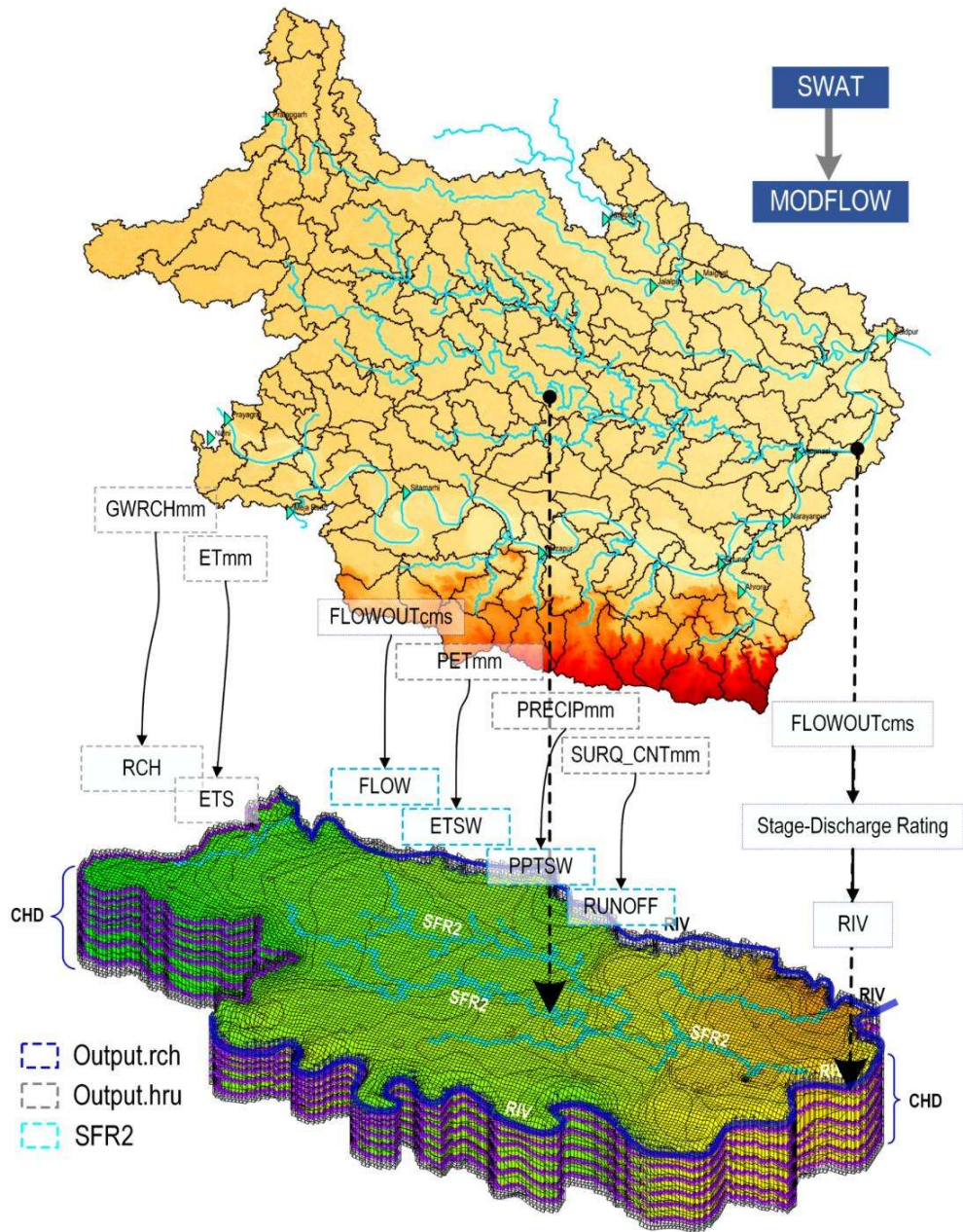


Figure 3.6. Illustration of the Integration of SWAT outputs to MODFLOW. The SWAT outputs and their hydrological variables mapped to the MODFLOW Packages (and their input variables)

The transient data outputs from the SWAT model are transferred to the MODFLOW-NWT packages through GUI support of GMS Software. The evapotranspiration and GW recharge data from the HRU output file is resampled to the subbasin area by taking the weighted average. The resampled subbasin scale data is mapped to the MODFLOW-

NWT Recharge (RCH) and Evapotranspiration (ETS) package. The SFR2 package gets its hydrological parameters from the combination of HRU outputs (Output.hru) and Reach outputs (Output.rch) of SWAT. The corresponding SWAT variables and MODFLOW-NWT packages have been illustrated in **Figure 3.6**. The river stages are calculated using statistical or ML models based on the established discharge rating curve from the observed data. This approach is beneficial in the case of forecasting models as the weather forecasts based on General Circulation Model outputs can be used with SWAT to simulate stream flow for future scenarios. The simulated flow can then be converted into the stage with observed data.

3.2.4 Calibration of GW model

Calibration of GW models is adjusting model parameters to ensure that the simulated results closely match observed data from the field, such as water levels, flow rates, or contaminant concentrations. This process is crucial because it enhances the reliability and predictive accuracy of the model. GW systems are inherently complex, with many uncertain variables like hydraulic conductivity, recharge rates, and boundary conditions (Mai, 2023). Proper calibration helps to reduce these uncertainties by fine-tuning parameters to reflect real-world conditions better (Anderson et al., 2015). Without accurate calibration, models could yield misleading results, potentially leading to poor management decisions, environmental degradation, or ineffective policy implementation (Mai, 2023). Calibration with discharge location data reduces differences in simulated discharge characteristics, improving predictions of GW discharge characteristics in river networks (Barclay et al., 2020; Rassam et al., 2013).

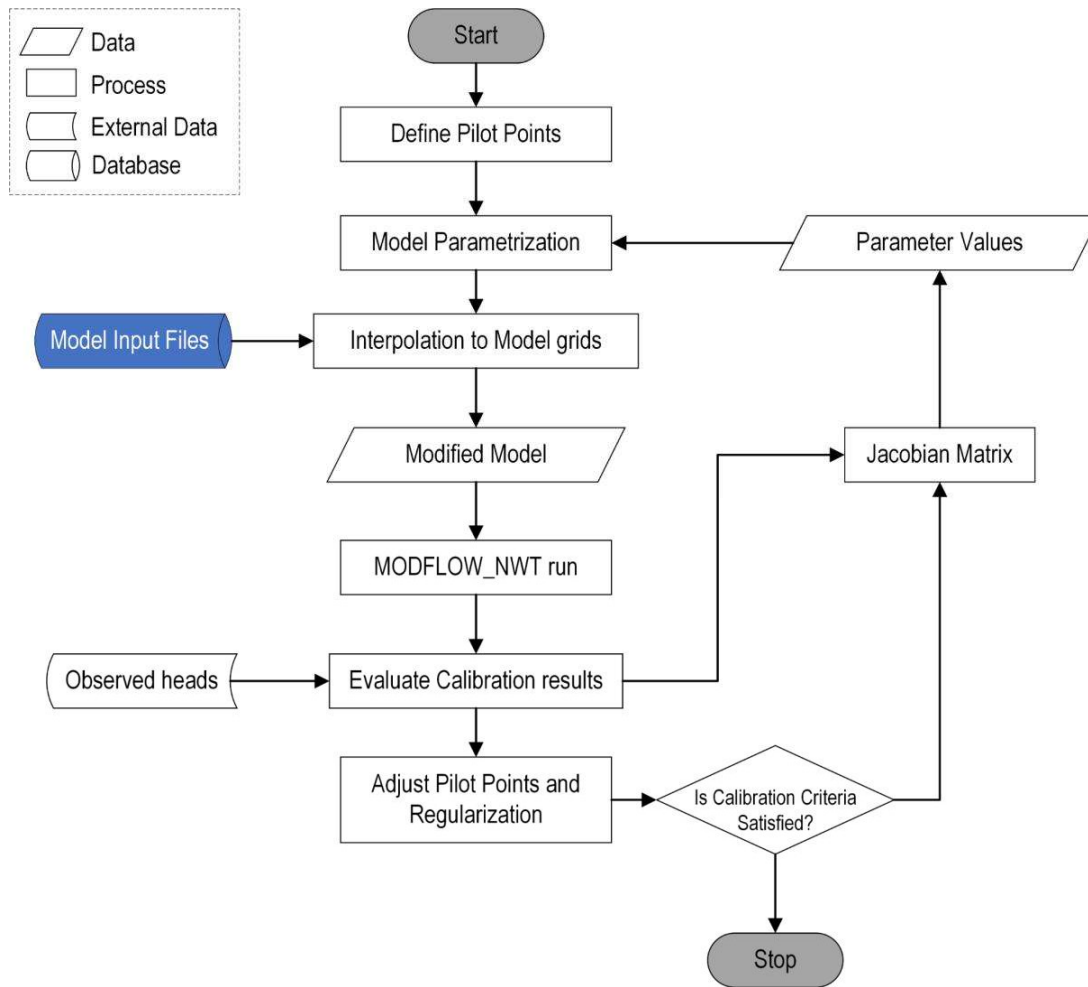


Figure 3.7. parameter calibration with pilot points in PEST

Parameter estimation (PEST) tool is a widely used approach for calibrating GW and SW models to observed data. To map the complex heterogeneity of aquifer parameters, the pilot point method (White et al., 2020) and regularization techniques have been employed. The pilot point method involves defining a set of points within the model domain, where the parameter values are treated as optimization variables during the calibration process. Regularization techniques such as Tikhonov regularization are used to constrain the parameter values and avoid overfitting. As the current version of GMS does not support the calibration of parameters in SFR2, a PSO optimization tool has been

developed to facilitate the calibration of the SFR2 parameters (Rafiei et al., 2022). The hydraulic conductivity of each segment of the river in the SFR2 package has been parameterized and calibrated such that in given hydrological inputs from the SWAT model, the SFR2 can better account for the baseflow to match the observed flow at different locations in the river. Calibrating the SFR2 parameters ensures reliable estimates of river aquifer exchanges. The overall flow chart for calibration of aquifer properties using pilot point method and later the PSO for SFR2 has been illustrated in **Figure 3.8**.

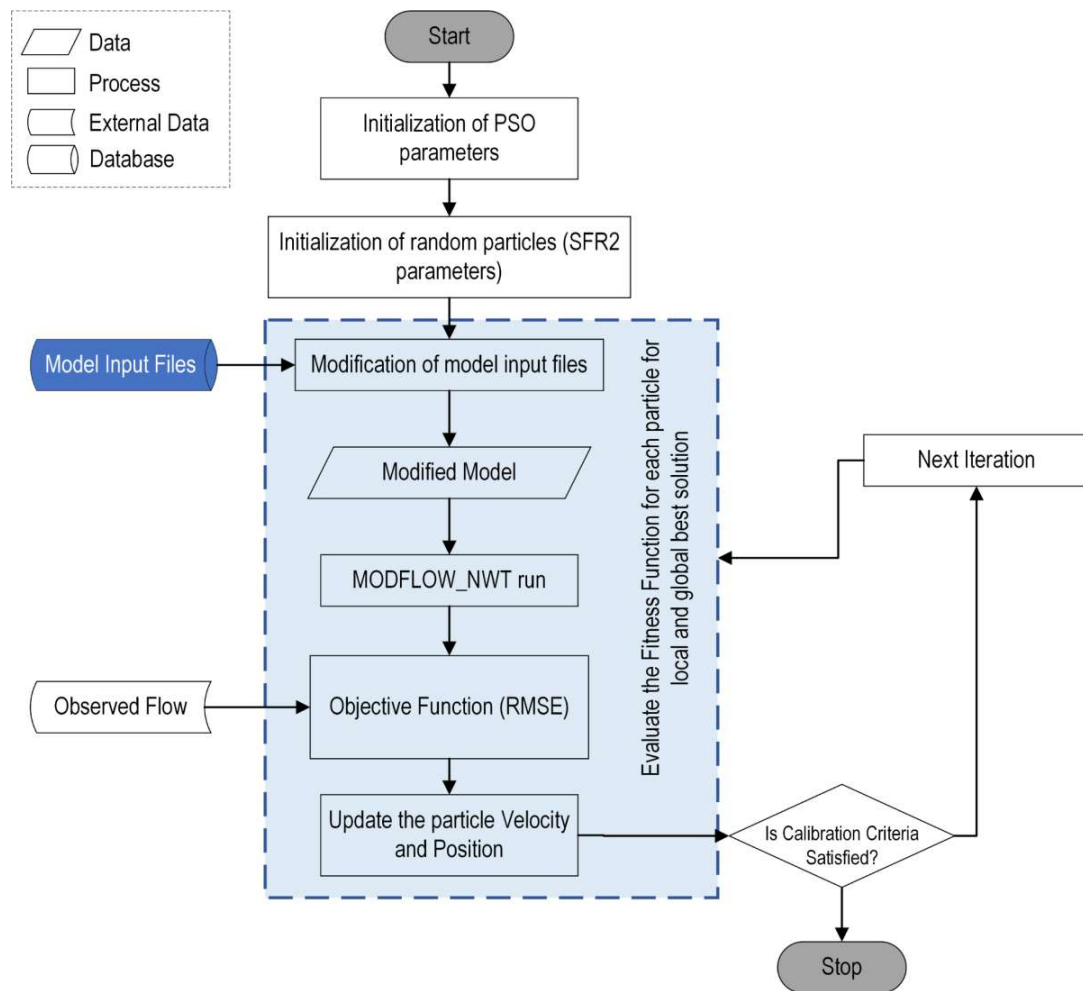


Figure 3.8. Parameter calibration of SFR2 with PSO

3.3 MACHINE LEARNING MODELS

Several ML models have been used for the preparation of thematic maps (such as land use), determining discharge-stage relationships, downscaling of coarse data, and forecasting the hydrological variables. The ML models used in this research work have been briefly explained below:

3.3.1 Artificial Neural Network (ANN)

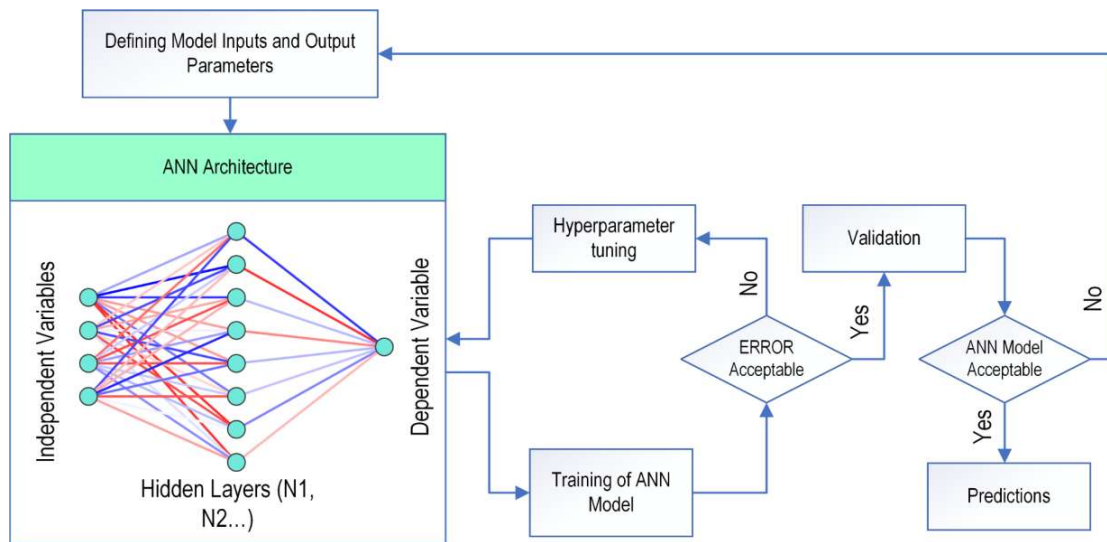


Figure 3.9. General ANN architecture and modeling process

Artificial neural networks (ANN) have been used in many research with good accuracy related to the field of GW potential zonation (Sameen et al., 2019; Tamiru and Wagari, 2022), GW simulation-optimization (Bahrami et al., 2016), Downscaling of GRACE-TWSA (Seyoum and Milewski, 2017) and Climate Model outputs (Nourani et al., 2018), and prediction of climate variables (Abhishek et al., 2012). The training data quality is associated with data management, including selecting the independent variables (predictors) having a high correlation with the dependent variable (the downscaled variable). The network architecture must be modified with optimal hyperparameters to

achieve a robust and accurate model for a given dataset (Vulpe-Grigoras and Grigore, 2021).

3.3.2 Random Forest

The RF models (Breiman, 2001) are widely used for hydrological data analysis and forecasting due to their ability to handle high-dimensional and complex data, account for nonlinear relationships, and identify important variables (Jing et al., 2020; Kumar et al., 2021; Sameen et al., 2019; Zabihi et al., 2016). Random Forest is an ensemble method that combines multiple decision trees to improve the predictive performance and reduce the risk of overfitting. In addition to low computational requirements, RF models have shown high prediction accuracy for data-scarce projects (Kumar et al., 2021).

The general modeling process using Random Forest (RF) models involves several key stages to ensure optimal performance. The dataset processing is required to clean the data for missing or outlier values. The independent variables are then selected or engineered based on the objective. The data is then split into training and testing data (generally into 30% for testing).

Following data preparation, the RF model is initialized. Key parameters in RF include the number of trees (`n_estimators`), the maximum depth of each tree (`max_depth`), and the minimum number of samples required to split a node (`min_samples_split`). Additionally, the feature subset size (`max_features`) controls the number of features considered at each split, contributing to the randomness and diversity of the trees. The output from each tree is aggregated, often by averaging in regression tasks, to generate the final prediction. These hyperparameters have been optimized using Bayesian optimization. The final tuned model is then used for prediction. The RF model has been built using the 'scikit

learn' library (Pedregosa et al., 2011) in Python. The 'scikit-optimize' library has been used to tune the hyperparameters of the RF model.

3.3.3 Long-Short Term Memory (LSTM)

Long Short-Term Memory (LSTM) models are a type of recurrent neural network (RNN) designed to handle sequential data and capture long-term dependencies, making them particularly suitable for time series prediction tasks. Key parameters of LSTM models include the number of layers, hidden state features, and input features, which are crucial for optimizing the model's performance. LSTMs are equipped with memory cells and gating mechanisms that allow them to retain information over extended periods, which is crucial for modeling complex temporal patterns. In hydrology, LSTMs have been effectively applied to tasks such as streamflow prediction, rainfall-runoff modeling, and hydrological variable forecasting. These models have demonstrated superior performance compared to traditional hydrological models, particularly in capturing the dynamics of hydrological processes and improving prediction accuracy in both gauged and ungauged basins (De La Fuente et al., 2024; Kratzert et al., 200AD; Lees et al., 2021; Liu et al., 2024).

The application of LSTM models in hydrology has led to significant advancements in streamflow estimation and hydrological forecasting. For instance, LSTMs have been used to develop hybrid models that integrate physically based hydrological models with deep learning techniques, enhancing the accuracy of streamflow predictions across various catchments (Liu et al., 2024). Additionally, LSTM models have been adapted to incorporate attention mechanisms and multitask learning, further improving their ability to predict multiple hydrological variables simultaneously and handle complex

interdependencies within the hydrological cycle (Dai et al., 2023; Yan et al., 2024). These innovations have not only improved the predictive performance of hydrological models but also provided insights into the underlying hydrological processes, making LSTMs a valuable tool in the field of hydrology.

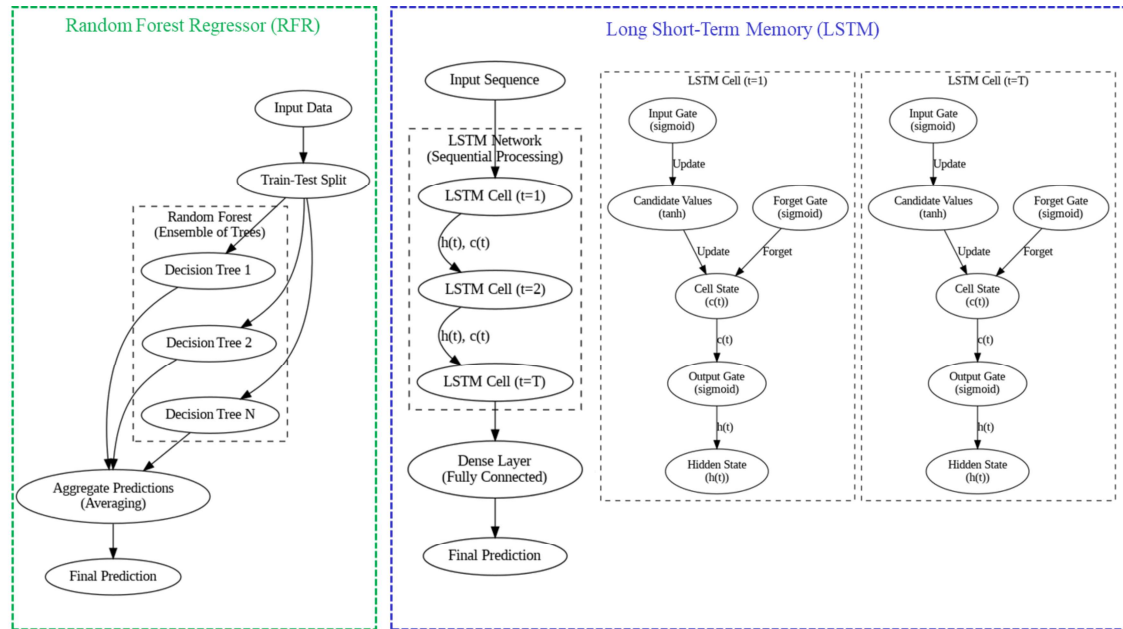


Figure 3.10. Sample architecture of Random Forest and LSTM Model

3.3.4 Hyperparameter tuning with Bayesian optimization

The model hyperparameters are an important factor in ensuring the model's performance and need to be tuned for the objective at hand. Instead of random search, the Bayesian optimization approach is the most robust and yields an optimum set of hyperparameters. Bayesian optimization uses a probabilistic model to guide the search for optimal hyperparameter values. It focuses on the most promising regions of the hyperparameter space by learning from previous iterations. The objective function (often model accuracy or error) is evaluated at each step. Bayesian optimization uses this information to propose new hyperparameter sets that are more likely to improve the model. This method is

particularly efficient when dealing with high-dimensional spaces and expensive evaluations, as it significantly reduces the number of iterations required to find optimal values.

3.4 ACCURACY MATRICES:

NSE and R^2 have been used to measure the model's accuracy. The *NSE* was used as the objective function during the calibration and validation of the SWAT Model. The ML Models were assessed for their prediction accuracy with *RMSE* and R^2 . Given that p_i is the predicted/simulated value corresponding to the actual/observed value o_i , \bar{o}_i is the actual/observed mean and \bar{p}_i is the mean of predicted/simulated values, the R^2 and *NSE* has been calculated as (Krause et al., 2005).

$$R^2 = \left[\frac{\sum_{i=1}^n (o_i - \bar{o}_i) (p_i - \bar{p}_i)}{\sqrt{\sum_{i=1}^n (o_i - \bar{o}_i)^2} \sqrt{\sum_{i=1}^n (p_i - \bar{p}_i)^2}} \right] \quad 3.12$$

$$NSE = 1 - \frac{\sum_i (o_i - p_i)^2}{\sum_i (o - \bar{o}_i)^2} \quad 3.13$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (o_i - p_i)^2 \quad 3.14$$

$$RMSE = \sqrt{MSE} \quad 3.15$$

3.5 STUDY AREA

The River Varuna is a tributary of the Ganga River, with its source located at Mailhan Jheel in the Phoolpur tahsil of Prayagraj District (coordinates: 25.599° N, 82.107° E). The river originates at an elevation of 112 meters and traverses a distance of approximately 224 kilometers before confluence with the Ganga River just northeast of Varanasi, where the elevation decreases to 65 meters. Consequently, the river exhibits a gentle gradient of

0.21 meters per kilometer. Covering a catchment area of roughly 3,664 square kilometers, as delineated by the SRTM Digital Elevation Model, the Varuna River is regarded as a medium-sized tributary of the Ganga River (Mishra et al., 2019). Throughout its course, it flows through the Prayagraj, Bhadohi, and Varanasi districts within Uttar Pradesh. The entire basin of the Varuna River is characterized by unconsolidated, recent flood alluvium (Raju et al., 2009).

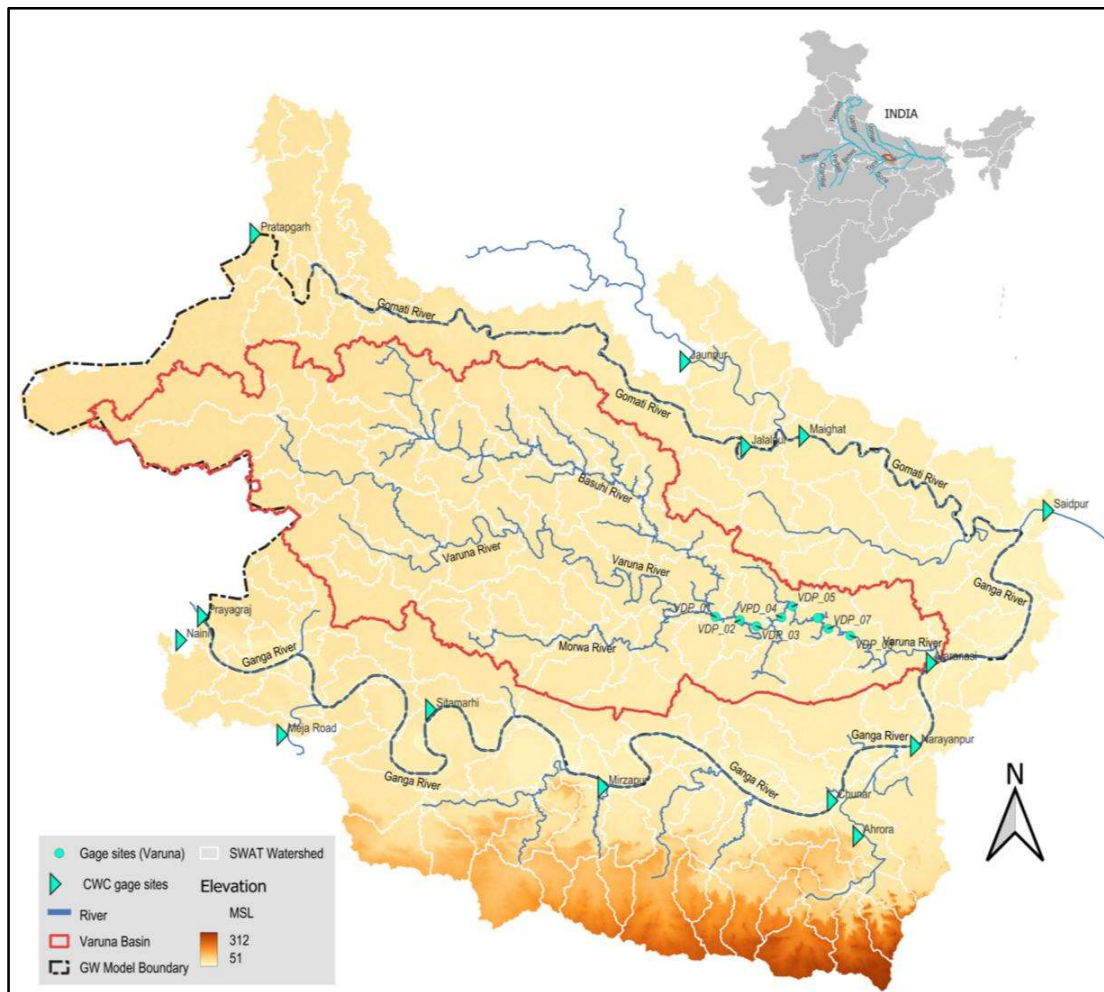


Figure 3.11. The study area and model domains

While the region predominantly displays a flat alluvial plain (Mishra et al., 2019), it also features sandy surfaces at relatively elevated positions (high-level floodplains) and clay-

rich surfaces at lower elevations (low-level floodplains). Similar to the broader Ganga River floodplain, this area is interspersed with numerous *tals* (shallow large depressions or basins), ponds, meander scrolls, and oxbow lakes. The region experiences an average annual rainfall of approximately 872.8 mm, as indicated by IMD 2D data from 1975 to 2020. It exhibits a temperature range from 4°C to 47°C, categorizing it within a semi-arid climatic zone.

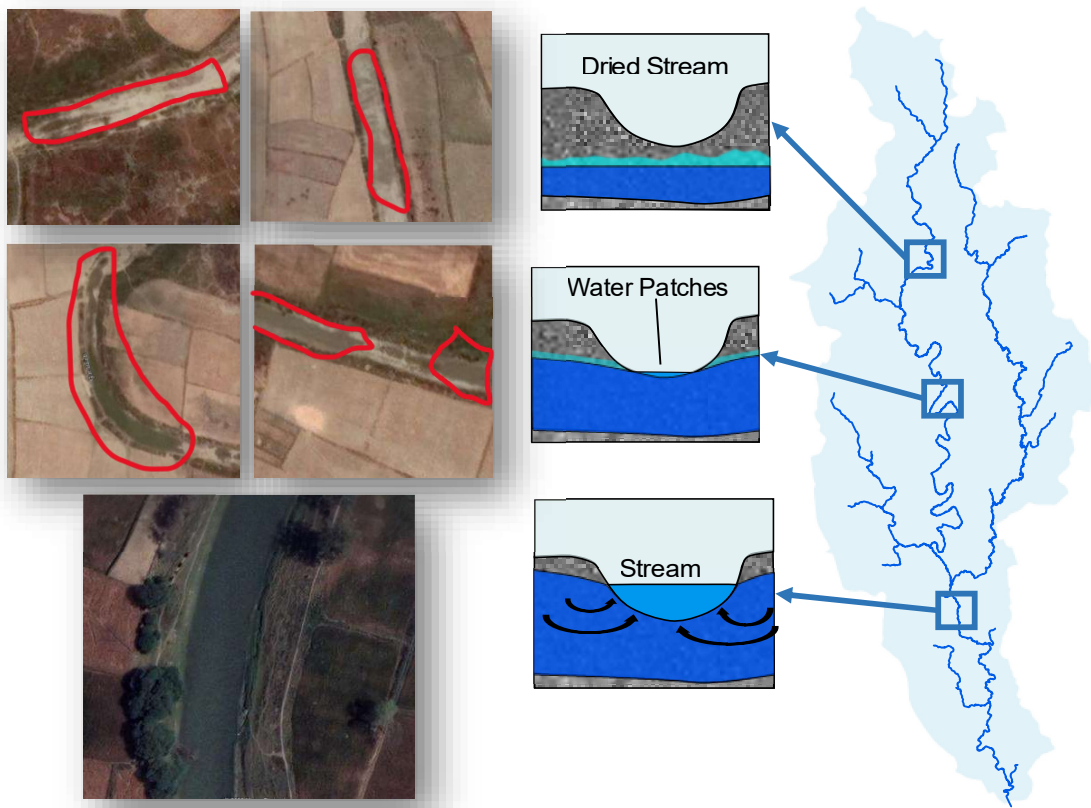


Figure 3.12. The State of RAE in the Varuna River Basin

The peak discharge in Varuna has shown variation from $\sim 40 \text{ m}^3/\text{sec}$ to $\sim 640 \text{ m}^3/\text{sec}$ with an average of $237.2 \text{ m}^3/\text{sec}$ at one of the gaging stations near Tendui in Varanasi. Climate change has resulted in increased peak discharge events, causing flooding in Varanasi. The

net discharge at the Tendui gage site varies from 1.2 m³/sec to 642 m³/sec with an average flow rate of 48.66 m³/sec. The Dry flow in the Varuna River is quite low (1.2 to 8.9 m³/sec) compared to the average flow rate. The river shows ephemeral flow conditions on the upstream reaches with patches of water (**Figure 3.12**). The flow downstream reaches near Varanasi city and is available throughout the year, with inflow from GW and drains.

3.6 FIELD VISITS AND DATA COLLECTION

Extensive data collection has been performed to satisfy the need to model the Varuna River Basin. The river discharge has been measured with an Acoustic Doppler Current Profiler (ADCP) (Sonotek M9) at various locations on the Varuna, Ganga, and Gomati Rivers. The river stages have also been measured using a Differential Global Positioning System (DGPS) (Sokkia GRX2). The resistivity data of underlying aquifers has been collected with a Schlumberger array using a resistivity meter. The saturated hydraulic conductivities along the basin have been collected with a Guelph permeameter and fused with the existing data from FAO (Food and Agriculture Organization of the United Nations).



Figure 3.13. Photos during data collection
