Finite Element Analysis

Alex Carey

December 20, 2025

1 Poisson Equation

System Setup

The Poisson system is defined by the following partial differential equation 1 with f defined on the domain Ω in 2 and boundary conditions defined on the boundary $\partial\Omega$ in 3.

$$-\nabla^2 u = f \quad \text{in } \Omega \tag{1}$$

$$f(x, y) = 2\pi^2 \sin(\pi x)\cos(\pi y) \tag{2}$$

$$u(0, y) = u(1, y) = 0, \quad \partial_y u(x, y)|_{y=0} = \partial_y u(x, y)|_{y=1} = 0 \tag{3}$$

the system admits the exact solution $u(x, y) = \sin(\pi x)\cos(\pi y)$ which we verify by subsitution into 1.

$$-\nabla^2(\sin(\pi x)\cos(\pi y)) = 2\pi^2 \sin(\pi x)\cos(\pi y) = f(x, y) \tag{4}$$

and further we consider the derivative with respect to y given $\partial_y u(x, y) = -\sin(\pi x)\sin(\pi y)$ which is zero at $y = 0$ and $y = 1$ as required by Neumann boundary condition. The Dirichlet boundary condition is also clearly satisfied.

Question 1

The Ritz-Galerkin principle states that the solution to the Poisson Equation is equivalent to the minimiser of the function defined

$$I[u] = \int_\Omega \left(\frac{1}{2}|\nabla u|^2 - fu\right) d\Omega. \tag{5}$$

To find this minimiser we consider the variation of $I[u]$ with respect to some variational function δu such that $u \to u + \delta u$ is considered. To maintain the boundary conditions for this new function we require that $\delta u = 0$ on the Dirichlet boundary and $\partial_y \delta u = 0$ on the Neumann boundary. We consider the varition in some limit $\epsilon \to 0$ such that

$$\frac{dI}{d\epsilon} = \int_\Omega \left(\frac{1}{2}\nabla u \cdot \nabla \delta u - f\delta u\right) d\Omega. \tag{6}$$

If u is the minimiser of I then this variation must be zero for all possible δu which gives us the weak form of the Poisson equation. We have that u is a solution to the Poisson system if u satisfies

$$\int_\Omega \nabla u \cdot \nabla \delta u \, d\Omega = \int_\Omega f \delta u \, d\Omega \quad \forall \delta u \tag{7}$$

where δu satisfies the altered boundary conditions described above.

The test function v that is used in the weak form if equivalent to the variational function δu as both functions are arbitrary and satisfy the same boundary conditions.

Question 2

We introduce the finite element basis functions $\Phi = \{\phi_i : i = 1, \ldots, N\}$ defined on the domain Ω such that we can approximate the solution

$$u \approx u_h = \sum_{j=1}^{N} u_j \phi_j \tag{8}$$

where u_j are the coefficients of the basis functions and N is the number of basis functions. We also consider the set of test functions to be equivalent to the set of basis functions such that $v = \phi_i$ for some $i = 1, \ldots, N$. As the basis functions span the solution space and are linearly independent, we only require that the weak form is satisfied for each basis function.

Substituting u_h into the Ritz-Galerkin formulation gives us

$$J[u_h] = \int_\Omega \left(\frac{1}{2} \left| \nabla u_h \right|^2 - f u_h \right) d\Omega. \tag{9}$$

where we aim to find the minimal coefficients u_j. We define the tensors $\mathbf{M}$ and $\mathbf{b}$ as follows

$$M_{ij} = \int_\Omega \nabla \phi_i \cdot \nabla \phi_j \, d\Omega \tag{10}$$

$$b_i = \int_\Omega f \phi_i \, d\Omega \tag{11}$$

such that the functional can be expressed as

$$J[\mathbf{u}] = \frac{1}{2} \mathbf{u}^T \mathbf{M} \mathbf{u} - \mathbf{u}^T \mathbf{b} \tag{12}$$

where $\mathbf{u} = [u_1, u_2, \ldots, u_N]^T$ is the vector of coefficients.

We consider also the substitution into the weak formulation giving

$$\int_\Omega \nabla \left(\sum_{j=1}^{N} u_j \phi_j \right) \cdot \nabla \phi_i \, d\Omega = \int_\Omega f \phi_i \, d\Omega \quad \forall i = 1, \ldots, N. \tag{13}$$

where we apply the same matrix operators to express this as

$$\mathbf{M} \mathbf{u} = \mathbf{b}. \tag{14}$$

We see that these two results are equivalent when we consider the minimsation of 12 with respect to $\mathbf{u}$ giving

$$\frac{dJ}{d\mathbf{u}} = \mathbf{M}\mathbf{u} - \mathbf{b} = 0 \tag{15}$$

recovering the weak form expression.

Question 3

We consider a quadrilateral mesh with the set of elements given by the tessellation of the domain

$$\Omega = \bigcup_{k=1}^{N_K} K_k \tag{16}$$

for $k = 1, \dots, N_K$ where K represent disjoint subsets of Ω such that and each is descibed by four nodes at the corners of the quadrilateral which we label $0, \dots, 3$ in a counter-clockwise manner starting from the bottom left corner.

We define a reference element $\hat{K} = [-1, 1] \times [-1, 1]$ with local coordinates $(\hat{x}, \hat{y})$ and define the map from the reference element onto a physical element K by some function $\mathbf{F}_K : \hat{K} \to K$ such that

$$\mathbf{x} = \mathbf{F}_K(\hat{x}, \hat{y}) = \sum_{\alpha=0}^{3} \mathbf{x}_{\alpha,k} \hat{\phi}_\alpha(\hat{x}, \hat{y}) \tag{17}$$

for some shape functions $\hat{\phi}_\alpha : \hat{K} \to [0, 1]$ defined on the element K_k. In the quadrilateral case we have four shape functions defined as follows

$$\hat{\phi}_0(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 - \hat{y}) \tag{18}$$

$$\hat{\phi}_1(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 - \hat{y}) \tag{19}$$

$$\hat{\phi}_2(\hat{x}, \hat{y}) = \frac{1}{4}(1 + \hat{x})(1 + \hat{y}) \tag{20}$$

$$\hat{\phi}_3(\hat{x}, \hat{y}) = \frac{1}{4}(1 - \hat{x})(1 + \hat{y}) \tag{21}$$

where $\mathbf{x}_\alpha$ are the physical coordinates of the nodes of element K.

The resulting Jacobian of this transformation is given

$$\mathbf{J}_K(\hat{x}, \hat{y}) = \begin{bmatrix} (1 - \hat{y})(x_{1,k} - x_{0,k}) + (1 + \hat{y})(x_{2,k} - x_{3,k}) & (1 - \hat{y})(y_{1,k} - y_{0,k}) + (1 + \hat{y})(y_{2,k} - y_{3,k}) \\ (1 - \hat{x})(x_{3,k} - x_{0,k}) + (1 + \hat{x})(x_{2,k} - x_{1,k}) & (1 - \hat{x})(y_{3,k} - y_{0,k}) + (1 + \hat{x})(y_{2,k} - y_{1,k}) \end{bmatrix} \tag{22}$$

where $(x_{\alpha,k}, y_{\alpha,k})$ are the physical coordinates of node α of element K_k.

We can then reform the matrix system from Question 2 in terms of the reference elements with the key subsitution

$$\int_{K_k} g(x, y)\, dK = \int_{\hat{K}} g(\mathbf{F}_K)\, |\det(\mathbf{J}_K)|\, d\hat{K} \tag{23}$$

yielding the matrix entries

$$M_{ij} = \sum_{k=1}^{N_K} \int_{\hat{K}} \nabla\phi_i(\mathbf{F}_K) \cdot \nabla\phi_j(\mathbf{F}_K) |\det(\mathbf{J}_K)| \, d\hat{K} \tag{24}$$

$$b_i = \sum_{k=1}^{N_K} \int_{\hat{K}} f(\mathbf{F}_K)\phi_i(\mathbf{F}_K) |\det(\mathbf{J}_K)| \, d\hat{K} \tag{25}$$

In the case of the matrix $\mathbf{M}$ we further consider the gradient terms which require the inverse of the Jacobian such that $\nabla\phi_i(\mathbf{F}_K) = \mathbf{J}_K^{-1}\hat{\nabla}\hat{\phi}_i$ where $\hat{\nabla}$ is the gradient operator with respect to the reference coordinates. This gives us the final form

$$M_{ij} = \sum_{k=1}^{N_K} \int_{\hat{K}} \left(\mathbf{J}_K^{-1}\hat{\nabla}\hat{\phi}_i\right) \cdot \left(\mathbf{J}_K^{-1}\hat{\nabla}\hat{\phi}_j\right) |\det(\mathbf{J}_K)| \, d\hat{K} \tag{26}$$

1.1 Code Output

A numerical implementation was created in python with the following outputs. The following figures show the output for u_h computed in Figure 1.The differences to the exact solution appear as in Figures 2, 3 and 4 for grid sizes of 128, 64 and 16 respectively.

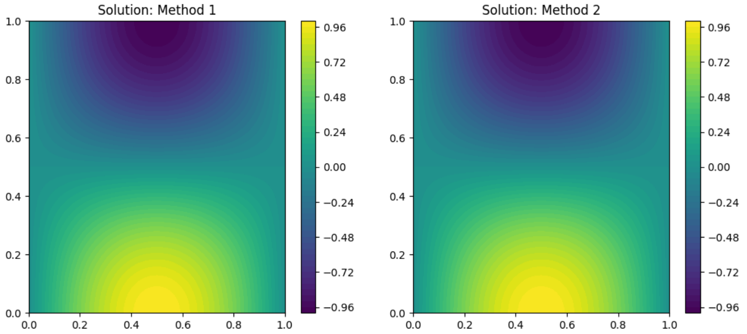

Figure 1: Finite Element Solution u_h to the Poisson Equation with a grid size of 128 and polynomials of order 1.

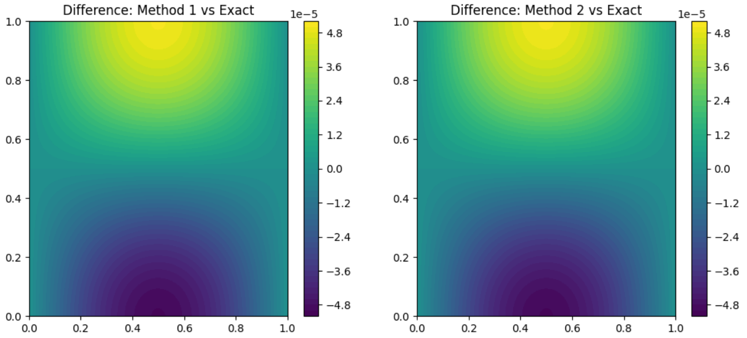

Figure 2: Error $u - u_h$ between the finite element solution and exact solution for a grid size of 128 and polynomials of order 1.

We have the L2 Errors for a number of different mesh sizes and polynomial orders as seen in Figure 5. We observe the order 1 convergence rate for the linear elements, order 2 for the quadratic elements and a

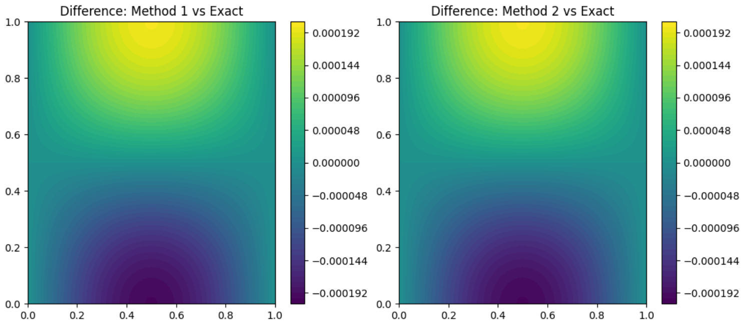

Figure 3: Error $u - u_h$ between the finite element solution and exact solution for a grid size of 64 and polynomials of order 1.

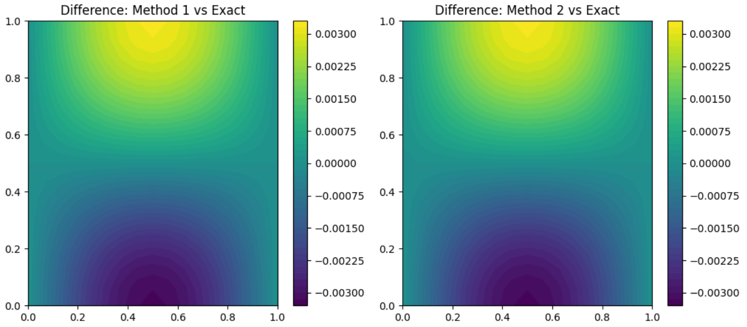

Figure 4: Error $u - u_h$ between the finite element solution and exact solution for a grid size of 16 and polynomials of order 1.

slightly higher order for the cubic elements. Note that this behaviour begins to degenerate for the higher orders as the round-off errors begin to dominate for large numbers of elements and high polynomial orders where more operations are required.

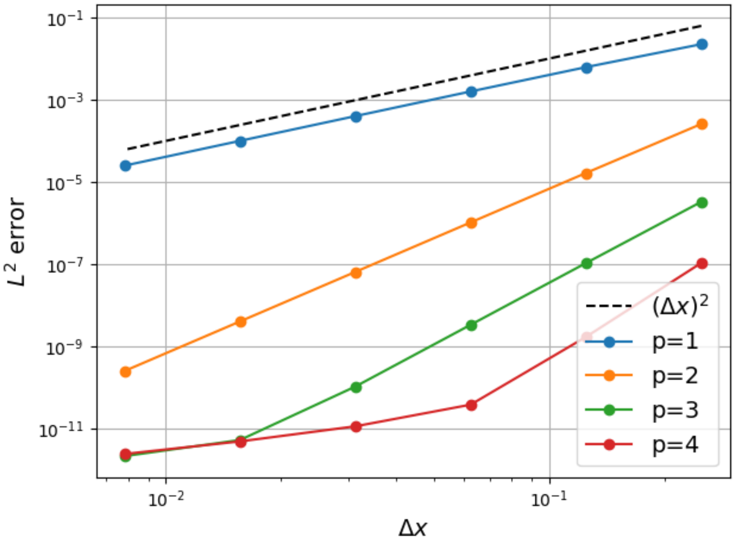

Figure 5: L2 Error convergence for different mesh sizes and polynomial orders.

For the alternate system formulation with the exact solution given

$$u(x, y) = \sin(5\pi x)y^2(2y - 3) \tag{27}$$

and the corresponding $f(x, y)$ defined by

$$f(x, y) = -(-\pi^2 y^2(2y - 3) + 6(2y - 1))\sin(5\pi x) \tag{28}$$

we observe the following solution and the error

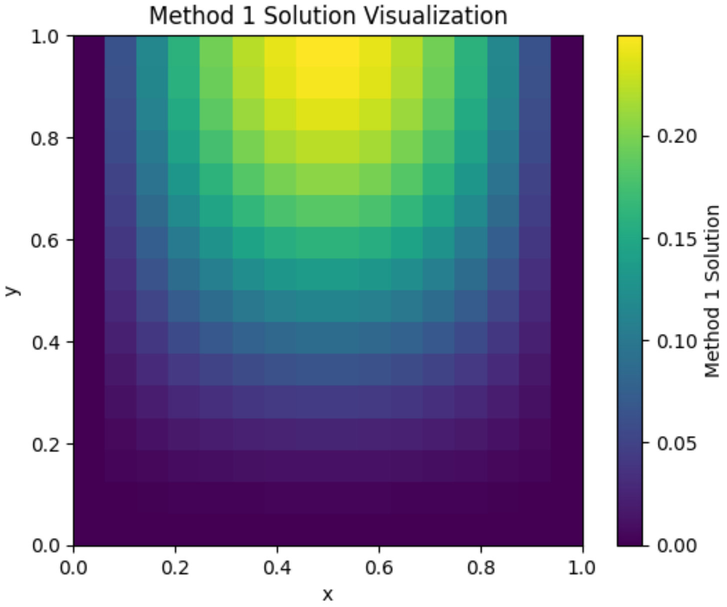

Figure 6: Finite Element Solution u_h to the alternate Poisson Equation .

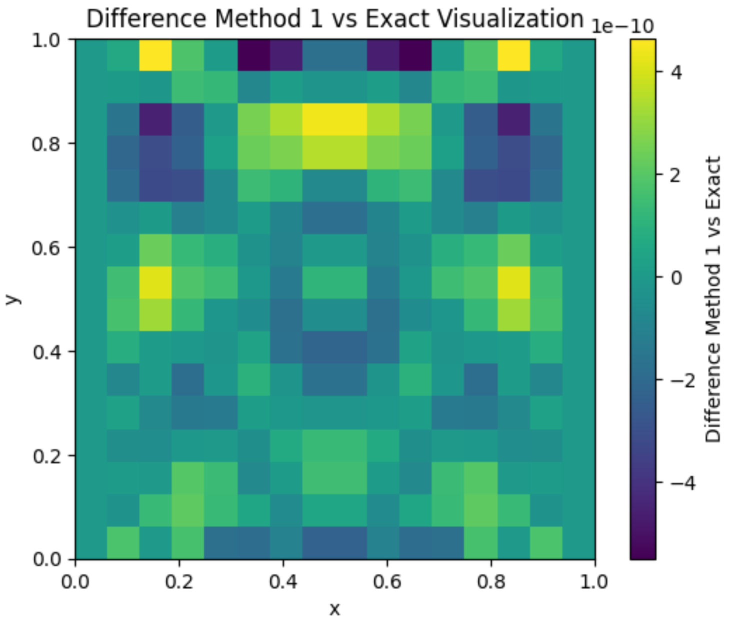

Figure 7: Error $u - u_h$ between the finite element solution and exact solution .

2 Groundwater and Canal Flow

System Setup

We consider the groundwater flow system defined by the following PDE with boundary conditions and ODE.

$$\partial_t(w_v h_m) - \alpha g \, \partial_y(w_v h_m \, \partial_y h_m) = w_v \frac{R}{m_{\text{por}} \sigma_e}, \qquad y \in [0, L_y], \tag{29}$$

$$\partial_y h_m = 0 \qquad \text{at } y = L_y, \tag{30}$$

$$h_m(0,t) = h_{cm}(t) \qquad \text{at } y = 0, \tag{31}$$

$$L_c w_v \frac{dh_{cm}}{dt} = w_v \frac{m_{\text{por}}}{2} \sigma_e \alpha g \partial_y \left(h_m^2\right)|_{y=0} - w_v \sqrt{g} \, \max\left(\frac{2}{3} h_{cm}(t), \, 0\right)^{3/2}. \tag{32}$$

for a groundwater system of in the region $[0, L_y]$ connected to a canal between $y = 0$ and $y = L_c$ where the height of the water is $h_m(y,t)$ in the groundwater and $h_{cm}(t)$ in the canal. The outflow flux from the canal is governed by a weir equation resulting in the equations above.

Question 1

We define the finite element basis function

$$\Psi = \{\phi_i : \phi_i \in C^0[0, L_y], \phi_i|_K \in \mathbf{P}_1(K)\} \tag{33}$$

for all elements K in the mesh of $[0, L_y]$ where $\mathbf{P}_1(K)$ is the space of linear polynomials on the element K. We approximate the solution as

$$h_m(y,t) \approx h_m^h(y,t) = \sum_{j=1}^{N} h_j(t)\phi_j(y) \tag{34}$$

where $h_j(t)$ are the time-dependent coefficients of the basis functions and N is the number of basis functions. We also consider the set of test functions to be equivalent to the set of basis functions such that $v = \phi_i$ for some $i = 1, \ldots, N$.

We define some test function $v(y,t)$ and multiply the PDE by this function to given

$$\int_0^{L_y} \partial_t(h_m^h)v\,dy - \int_0^{L_y} \alpha g\,\partial_y(h_m^h\,\partial_y h_m^h)v\,dy = \int_0^{L_y} \frac{R}{m_{\text{por}}\sigma_e} v\,dy. \tag{35}$$

For ease we define $F = \frac{1}{2}\alpha g \partial_y (h_m^h)^2$ and integration by parts gives

$$\int_0^{L_y} \partial_t(h_m^h)v\,dy = [Fv]_0^{L_y} - \int_0^{L_y} F\,\partial_y v\,dy + \int_0^{L_y} \frac{R}{m_{\text{por}}\sigma_e} v\,dy. \tag{36}$$

where Fv is representative of the flux into the canal. The Neumann boundary condition at $y = L_y$ gives $F(L_y) = 0$ and so we can eliminate the flux through this boundary by considering the ODE for the canal height. This results in the form

$$\int_0^{L_y} \partial_t(h_m^h)v\,dy + \frac{v_0 L_c \partial_t h_m(0,t)}{m_{\text{por}}\sigma_e} = -\int_0^{L_y} F\,\partial_y v + \frac{vR}{m_{\text{por}}\sigma_e} v\,dy. - \frac{v_0 Q_c}{m_{\text{por}}\sigma_e} \tag{37}$$

where the discharge $Q_c(h_m(0,t)) = \sqrt{g}\max\left(\frac{2}{3}h_m(0,t), 0\right)^{3/2}$.

Application of the explicit Euler method for discretisation of the time derivative yields

$$\int_0^{L_y} h_m^{n+1} v\,dy + \frac{v_0 L_c h_m^{n+1}(0)}{m_{\text{por}}\sigma_e} = \int_0^{L_y} h_m^n v\,dy + \frac{v_0 L_c h_m^n(0)}{m_{\text{por}}\sigma_e} - \Delta t\left(\int_0^{L_y} F^n\,\partial_y v + \frac{vR^n}{m_{\text{por}}\sigma_e} v\,dy - \frac{v_0 Q_c^n}{m_{\text{por}}\sigma_e}\right) \tag{38}$$

We define the tensors as follows

$$M_{ij} = \int_0^{L_y} \phi_i \phi_j\,dy \tag{39}$$

$$b_i^n = -\int_0^{L_y} \alpha g h_m^n\,\partial_y \phi_i + \frac{R^n}{m_{\text{por}}\sigma_e}\phi_i\,dy \tag{40}$$

such that the scheme can be expressed in matrix form as

$$M_{ij}h_j^{n+1} + \frac{\delta_{i1}L_c}{m_{\text{por}}\sigma_e}h_m^{n+1}(0) = M_{ij}h_j^n + \frac{\delta_{i1}L_c}{m_{\text{por}}\sigma_e}h_m^n(0) + \Delta t\left(b_i^n - \frac{v_0 Q_c^n}{m_{\text{por}}\sigma_e}\right) \tag{41}$$

where we have used the Kronecker delta δ_{i1} to select the test function at the boundary $y = 0$ (all other test functions are zero at this point).

Code Output

$\theta = 0$, $\mathbf{P_1}$(Linear Elements)

A constant rainfall case with a fully explicit scheme on linear elements.

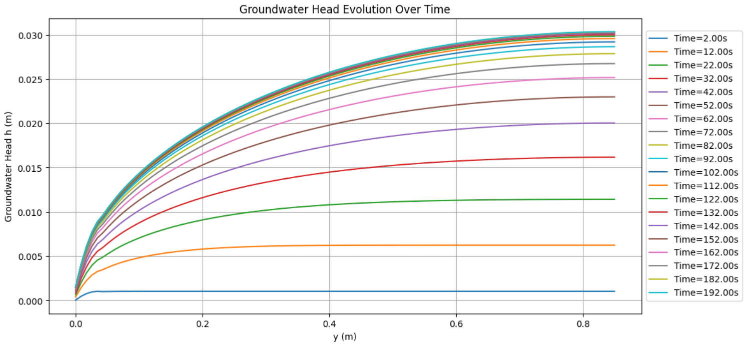

Figure 8: Plot of the groundwater height h_m at various times.

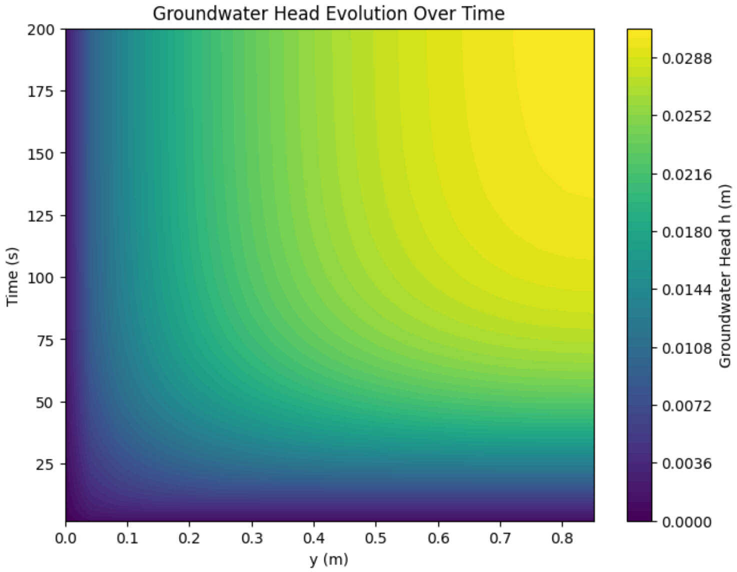

Figure 9: Contour plot of the groundwater height h_m over time.

$\theta = 0.5$, $\mathbf{P_2}$(Quadratic Elements)

A constant rainfall case with a Crank-Nicolson scheme on quadratic elements.

Variable Rainfall: $\theta = 0.5$, $\mathbf{P_2}$(Quadratic Elements)

A variable rainfall case with a Crank-Nicolson scheme on quadratic elements.

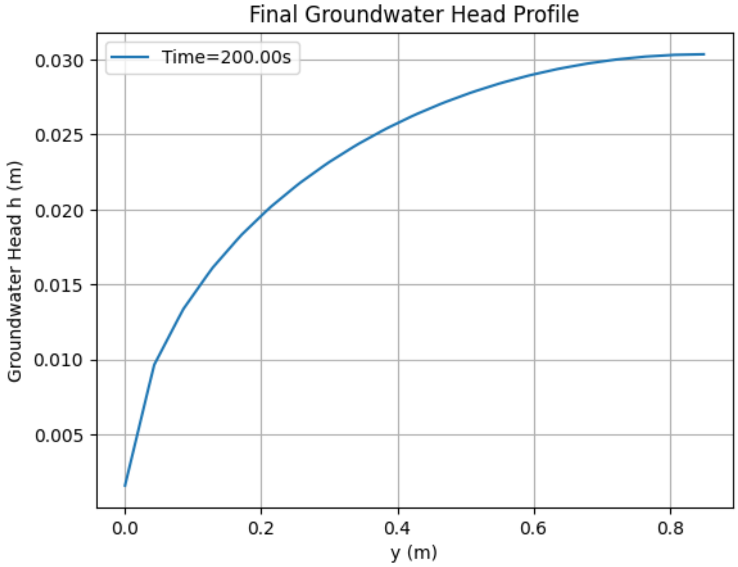

Figure 10: Plot of the final (steady) groundwater profile..

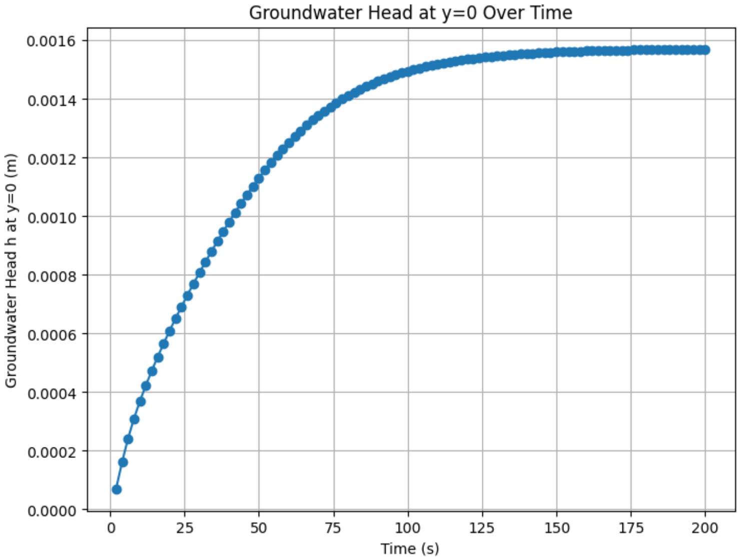

Figure 11: Plot of the canal height h_{cm} over time.

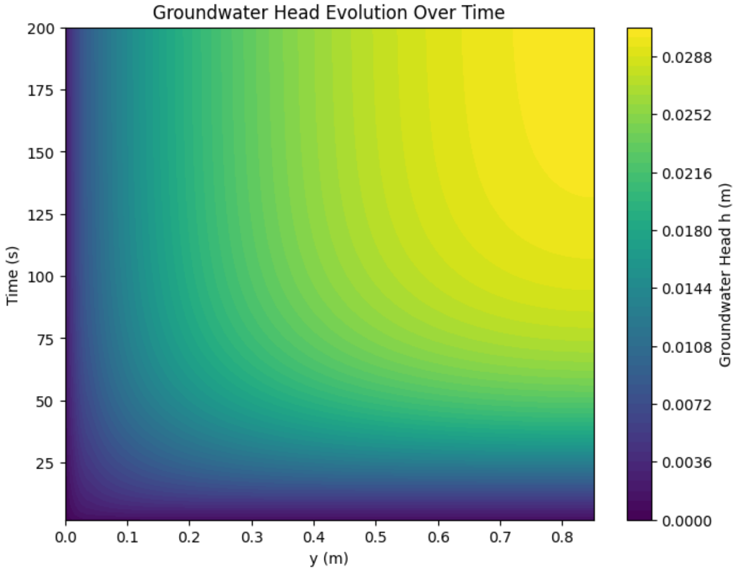

Figure 12: Contour plot of the groundwater height h_m over time.

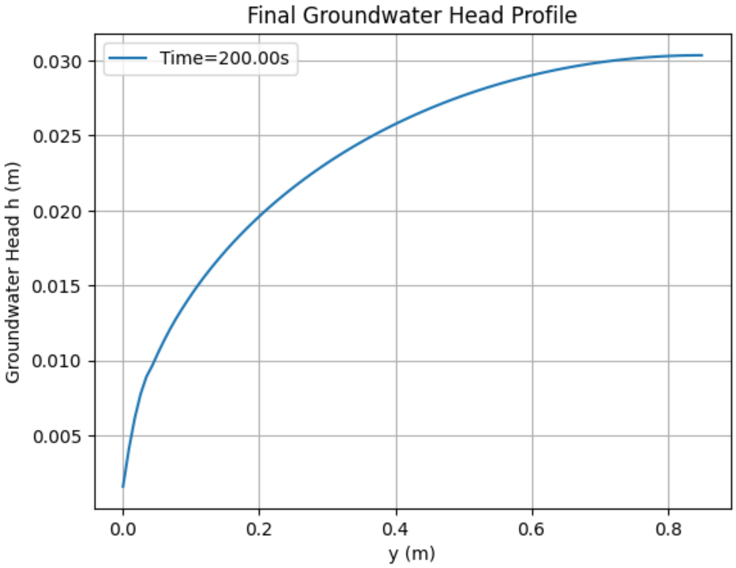

Figure 13: Plot of the final (steady) groundwater profile..

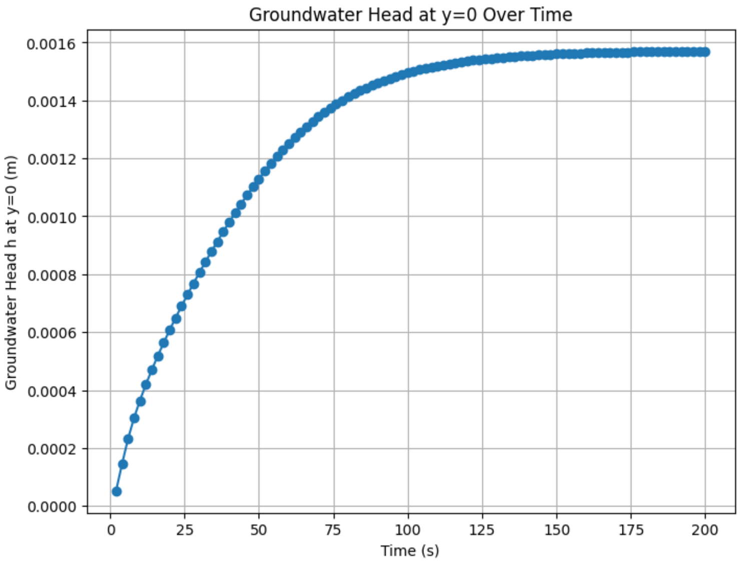

Figure 14: Plot of the canal height h_{cm} over time.

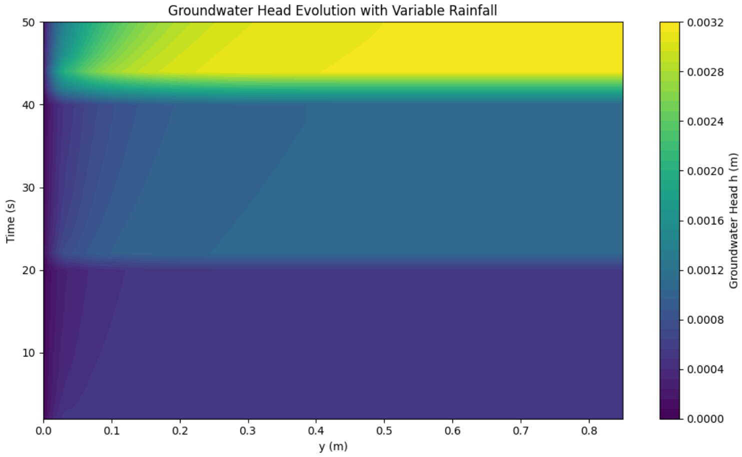

Figure 15: Contour plot of the groundwater height h_m over time.

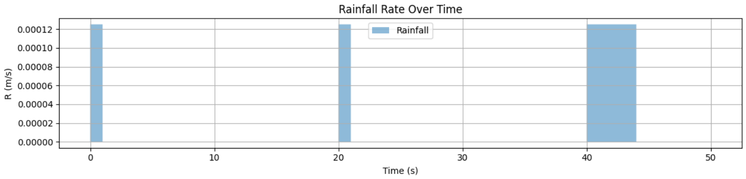

Figure 16: Plot of the activation of rainfall over time.

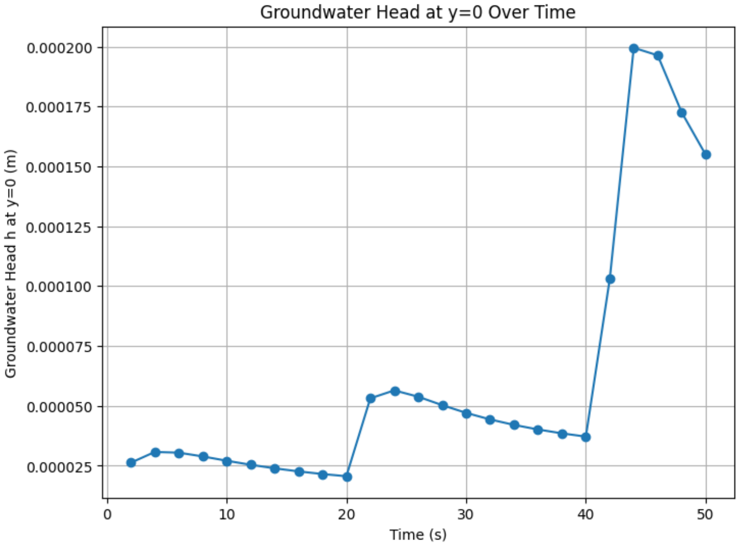

Figure 17: Plot of the canal height h_{cm} over time.