Electronic Transactions on Numerical Analysis.
Volume 65, pp. 553–578, 2026.
© 2026, the Author(s), licensed via CC BY 4.0.
ISSN 1068–9613.
DOI: 10.1553/etna_vol65s553

ETNA
Kent State University and
Johann Radon Institute (RICAM)

AN EXACT PIECEWISE AFFINE SOLVER FOR THE DISCRETIZED ENTHALPY FORMULATION OF THE STEFAN PROBLEM IN PERMAFROST MODELING*

DAVID HÖTTEN†, JENNY NIEBSCH‡, RONNY RAMLAU‡, AND WALTER ZULEHNER‡

Abstract. We consider the numerical solution of systems of equations arising from the discretization of the weak formulation of the Stefan problem. In permafrost modeling within land surface models and dynamic global vegetation models, solving these systems requires low computational cost and an extremely robust solver to enable large-scale, long-term simulations. Instead of applying a Gauss–Seidel-type method, we leverage the piecewise affine structure of the system and solve it using the Katzenelson algorithm. We prove that, up to floating-point errors, the exact solution is found in a finite number of steps from almost every initial guess. Comparisons with the Neumann analytical solution and tests in the Lund-Potsdam-Jena managed land (LPJmL) vegetation model show that the new method does not introduce significantly higher computational costs than the widely used decoupled energy-conservation parameterization (DECP) method while providing greater accuracy.

Key words. Stefan problem, enthalpy formulation, permafrost, decoupled energy-conservation parameterization, piecewise affine function, Katzenelson algorithm

AMS subject classifications. 65M60, 65M22

1. Introduction.

Human activities have caused a transformation of all parts of the Earth system, including the atmosphere, land, ocean, and cryosphere. The need to discern and evaluate the past and potential future impacts of humans on their environment makes process-based computer models indispensable tools, since they allow for the consideration of counterfactual scenarios. Dynamic global vegetation models (DGVMs) and land surface models (LSMs) represent the biophysical water, energy, and substance cycles that occur within land systems. Because these dynamics are often nonlinear and the models operate on global grids, highly efficient numerical solvers are required. An example is soil heat conduction in permafrost regions, which cover up to 22% of the northern hemisphere [15]. Capturing the seasonal freeze-thaw cycles in these soils requires accounting for the release and absorption of latent heat. This creates nonlinear and nonsmooth dynamics that standard solvers struggle to handle efficiently.

The numerical solution of the heat equation with phase change, commonly referred to as the Stefan problem, is a well-established and extensively studied subject in numerical analysis [4, 5, 22]. The algebraic system of equations arising from the discretized enthalpy formulation is often solved with a Gauss–Seidel-type method [12, 23]. A prominent example in the latter class of methods was developed and analyzed by White [26, 27]. The starting point of this method is a backward Euler finite-difference discretization of the enthalpy formulation. White then proves convergence to the discrete solution for his Gauss–Seidel-type algorithm, provided the time step is sufficiently small and the solution from the previous time step is used as the initial guess. More precisely, two upper bounds on the time step are required for the proof: the first is given explicitly in terms of the data, whereas the second is only shown to exist and cannot be determined explicitly. Furthermore, this framework allows the thermal conductivity and the temperature–enthalpy relation to depend on space and on the solution, with potentially nonlinear dependence on the solution.

A different strategy is the decoupled energy-conservation parameterization (DECP), a term introduced by Zhang et al. [28], which is widely used in DGVMs and LSMs such

*Received March 24, 2025. Accepted September 9, 2026. Published online on October 29, 2026. Recommended by Herbert Egger.
†Corresponding author. Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, P.O. Box 6012 03, D-14412 Potsdam, Germany (davidho@pik-potsdam.de).
‡Radon Institute for Computational and Applied Mathematics, Altenbergerstraße 69, 4040 Linz, Austria.
This work is published by ETNA and licensed under the Creative Commons license CC BY 4.0.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

as the community land model 5 [11] and the community Noah land surface model with multiparameterization options [14]. In DECP, the heat-conduction step is performed using the standard heat equation, and the effects of phase change are subsequently incorporated through a correction based on energy conservation. The heat-conduction step is typically solved with an implicit scheme such as the Crank–Nicolson method [11, 25]. By decoupling heat conduction and phase change, DECP avoids solving a nonlinear system at each time step. However, the method is known to produce an artificial stretching of the freezing region, which can lead to inaccuracies [13]. A comprehensive overview of the governing equations and discretization schemes used in different land models is provided in [20, table 1].

The application of an enthalpy-based scheme in DGVMs and LSMs presents two central challenges. First, the numerical method has to be robust enough to solve the equations for approximately 60,000 soil columns with heterogeneous forcing and parameters in parallel across simulation periods exceeding 4,000 years without convergence failures. Second, the computational cost should remain sufficiently low for simulations at this scale. These requirements put limitations on the use of White's method. The fact that one of the time-step restrictions is not known explicitly means that convergence cannot be verified for a given problem. Although White's numerical experiments suggest that the second bound may not be restrictive in practice, the lack of theoretical convergence guarantees is problematic in light of the robustness requirements. Moreover, the reported number of iterations may already render the method too computationally expensive for the intended application. Conversely, while DECP is computationally attractive and avoids nonlinear iterations altogether, it does so at the cost of introducing additional approximation errors.

This study develops and evaluates an alternative to White's iterative enthalpy solver [27] and the DECP approach. Like White's method, the proposed approach finds a solution to the algebraic system of equations arising from a discretization of the enthalpy formulation of the Stefan problem. Indeed, the discretization considered here is very similar to the one in [27]. However, the algorithm used to solve the resulting system of equations is fundamentally different. By dropping the full nonlinearity of the constitutive equations and assuming piecewise affine and piecewise constant relations—as is standard in permafrost modeling—we show that the system itself possesses a piecewise affine structure. This structural property is then exploited by applying the Katzenelson algorithm [9], which can be viewed as a Newton-type iteration with a specific damping strategy or a piecewise linear homotopy method [18]. Utilizing the theory of piecewise affine functions and the properties of the matrices involved, we show that the algorithm converges to the exact solution of the discrete problem in a finite number of iterations independent of the time step.

For readers who are not familiar with the Stefan problem and its enthalpy formulation, we recommend the introductory texts [2] and [21]. An introduction to the finite-element method, which underlies the derivation of the proposed scheme, can be found in [29] and [30]. The theory of piecewise affine functions is introduced in [16].

The remainder of this paper is organized as follows. In Section 2, we introduce the enthalpy formulation of the Stefan problem and a finite-element discretization represented as a nonlinear algebraic system of equations; the derivation of this discretization is presented in Appendix A. In Section 3, we prove that solutions to the system of equations are roots of a piecewise affine homeomorphism by employing the theory of piecewise affine functions. Following [7], we then demonstrate the convergence of the Katzenelson algorithm for the given problem under simplifying assumptions. Section 4 presents the results from a mesh-convergence study and from soil-temperature simulations in the Lund-Potsdam-Jena managed land (LPJmL) vegetation model using the proposed method. Finally, Section 5 compares the new method with White's method and the DECP approach.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

2. The discrete problem. Land models like LPJmL often compute output variables on a spatial grid with a resolution of 0.5° in both latitude and longitude. The soil temperature within each grid cell is assumed to vary only with depth, while remaining uniform in the horizontal directions. Consequently, heat conduction is modeled as a one-dimensional process in space. The space-time domain is defined as

$$G := (0, D) \times (0, T),$$

where D denotes the maximum depth of the soil column and T represents the total simulation duration. On this domain, the given spatially dependent data include the frozen and unfrozen volumetric heat capacities, $c_f(x)$ and $c_u(x)$, respectively, as well as the frozen and unfrozen thermal conductivities, $k_f(x)$ and $k_u(x)$. Additionally, $L(x)$ represents the volumetric latent heat of fusion.

Heat conduction with phase change can be described either in terms of the temperature alone or in terms of temperature and enthalpy. The classical description is the two-phase Stefan problem [2]. Its unknowns are the temperature $u(x,t)$ and the trajectory $X : [0,T] \to (0,D)$ of a single sharp interface separating the frozen from the unfrozen part of the soil column. In the frozen region $x < X(t)$ the standard linear heat equation $c_f u_t = (k_f u_x)_x$ holds; in the unfrozen region $x > X(t)$ the same equation holds with c_u and k_u. The temperature at the interface is at the melting point, $u(X(t),t) = 0$. The interface trajectory is determined by an energy balance at the free boundary: the latent heat absorbed or released by the moving front equals the jump of the heat flux across it, which is expressed by the *Stefan condition*

$$L(X(t))X(t)_t = k_f(X(t))u(X(t)^-,t)_x - k_u(X(t))u(X(t)^+,t)_x, \tag{2.1}$$

with one-sided spatial derivatives on the right-hand side. Together with initial and boundary conditions, this closes the system. A limitation of this formulation is that it postulates a single interface. This renders it inadequate for permafrost soils, where one or more unfrozen ground layers may be entirely enclosed within an otherwise frozen soil column—a phenomenon known as *closed talik*—and where interfaces form and vanish over the seasonal freeze-thaw cycle.

The *enthalpy formulation* of heat conduction with phase change overcomes this limitation by introducing the volumetric enthalpy $e(x,t)$ as a second unknown alongside the temperature $u(x,t)$. The enthalpy tracks the local energy content, and thereby the phase state, at every point, so that no interface needs to be prescribed; phase boundaries emerge implicitly from the solution and their number may change over time. In scenarios where only a single interface is present, the enthalpy formulation is equivalent to the Stefan problem. In particular, the Stefan condition (2.1) can be recovered from it [2]. The temperature and the enthalpy are related by

$$u = \beta(e,x) := \begin{cases} e/c_f(x) & \text{for } e \le 0 & \text{(frozen)},\\ 0 & \text{for } 0 < e < L(x) & \text{(mushy)},\\ [e - L(x)]/c_u(x) & \text{for } e \ge L(x) & \text{(unfrozen)}. \end{cases} \tag{2.2}$$

Defining the thermal conductivity as a piecewise function of temperature,

$$k(u,x) := \begin{cases} k_f(x) & \text{for } u \le 0,\\ k_u(x) & \text{for } u > 0, \end{cases}$$

the mathematical model is the enthalpy formulation of heat conduction with phase change,

$$\begin{aligned} &e_t = [k(u,x)u_x]_x, \quad u = \beta(e,x) && \text{in } G,\\ &u(0,t) = s(t), \qquad u(D,t)_x = 0 && (0 \le t \le T),\\ &e(x,0) = e_0(x) && (0 < x < D), \end{aligned} \tag{2.3}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

with a prescribed surface temperature $s(t)$, a zero-flux Neumann condition at the bottom, and a given initial enthalpy $e_0(x)$. The system (2.3) is understood in the weak sense; the weak formulation is stated explicitly in Appendix A. The existence and uniqueness of solutions to the weak enthalpy formulation, for spatially dependent thermal conductivity but constant latent heat and heat capacity, have been proven in [6].

A standard finite-element discretization of (2.3) is given as follows; see Appendix A for its derivation. Consider a partition of the interval $(0, D)$ defined by the nodes $0 = x_0 < x_1 < \cdots < x_\kappa = D$, where $\kappa \in \mathbb{N}$. For $i = 1, \ldots, \kappa$, the subintervals $T_i := (x_{i-1}, x_i)$ are referred to as *elements* and the quantities $h_i := x_i - x_{i-1}$ are termed the *element sizes*. The time interval is discretized by $0 = t_0 < t_1 < \cdots < t_N = T$ and we use the notation $\Delta t_n := t_n - t_{n-1}$ for $n = 1, \ldots, N$ and $N \in \mathbb{N}$.

The following *data* enter the scheme. The thermal conductivities are assumed to be constant on each element. For each $i = 1, \ldots, \kappa$, the values $k_i^u, k_i^f \in \mathbb{R}$ denote the unfrozen and frozen thermal conductivities on the element T_i. We further define

$$k_i(u) := \begin{cases} k_i^f & \text{for } u \le 0, \\ k_i^u & \text{for } u > 0. \end{cases} \tag{2.4}$$

The heat capacities and the latent heat enter the scheme through their nodal values, $c_i^f := c^f(x_i)$, $c_i^u := c^u(x_i)$, and $L_i := L(x_i) \in \mathbb{R}$ for $i = 1, \ldots, \kappa$. We assume that

$$k_i^f, k_i^u, c_i^f, c_i^u, L_i > 0 \quad \text{for all } i \in 1, \ldots, \kappa. \tag{2.5}$$

The surface temperature at time t_n is given as $s^n \in \mathbb{R}$ for all $n = 0, \ldots, N$.

The unknowns of the discrete problem at time t_n are the vectors $\eta^n \in \mathbb{R}^\kappa$ and $\gamma^n \in \mathbb{R}^\kappa$ whose entries η_i^n and γ_i^n approximate the enthalpy and the temperature at the node x_i and time t_n. They are coupled by imposing the constitutive relation (2.2) at the nodes $x_1, \ldots, x_\kappa$. That is, defining the map

$$\mathcal{B} : \mathbb{R}^\kappa \longrightarrow \mathbb{R}^\kappa, \quad \mathcal{B}(\eta) := [\beta(\eta_1, x_1) \;\; \beta(\eta_2, x_2) \;\; \ldots \;\; \beta(\eta_\kappa, x_\kappa)]^{\mathrm{T}},$$

the two vectors satisfy

$$\gamma^n = \mathcal{B}(\eta^n).$$

We define

$$\begin{aligned} Q_1^n(\gamma^n) &:= \frac{k_1(\gamma_1^n)\gamma_1^n - k_1(s^n)s^n}{h_1}, \\ Q_j^n(\gamma^n) &:= \frac{k_j(\gamma_j^n)\gamma_j^n - k_j(\gamma_{j-1}^n)\gamma_{j-1}^n}{h_j} \quad \text{for } j = 2, \ldots, \kappa, \end{aligned} \tag{2.6}$$

and

$$F_i^n(\gamma^n) := \begin{cases} Q_i^n(\gamma^n) - Q_{i+1}^n(\gamma^n) & \text{for } i < \kappa, \\ Q_i^n(\gamma^n) & \text{for } i = \kappa, \end{cases} \tag{2.7}$$

for $i = 1, \ldots, \kappa$. The case $i = \kappa$ realizes the zero-flux boundary condition at the bottom. For each $n = 0, \ldots, N$, we obtain the map

$$F^n : \mathbb{R}^\kappa \longrightarrow \mathbb{R}^\kappa, \quad F^n(\gamma) := [F_1^n(\gamma) \;\; \ldots \;\; F_\kappa^n(\gamma)]^{\mathrm{T}}, \tag{2.8}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

which depends on time only through the boundary datum s^n entering Q_1^n. The lumped *mass matrix* is given by

$$M_h := \operatorname{diag}\big(\tfrac{1}{2}(h_1 + h_2),\ \ldots,\ \tfrac{1}{2}(h_{\kappa-1} + h_\kappa),\ \tfrac{1}{2}h_\kappa\big) \in \mathbb{R}^{\kappa\times\kappa}.$$

Now we can state the discretized version of (2.3). Let $\theta \in [0,1]$, where the cases $\theta = 0$, $\theta = 1$, and $\theta = \frac{1}{2}$ correspond to the forward Euler, backward Euler, and Crank–Nicolson time discretizations, respectively. Let $n = 0, \ldots, N-1$. The time stepping is initialized with the nodal values of the initial enthalpy $\eta_i^0 := e_0(x_i)$ for $i = 1, \ldots, \kappa$.

Discrete problem in algebraic form. Given $\eta^n \in \mathbb{R}^\kappa$, find $\eta^{n+1} \in \mathbb{R}^\kappa$ such that

$$\frac{1}{\Delta t_{n+1}} M_h(\eta^{n+1} - \eta^n) = -(1-\theta)[F^n \circ \mathcal{B}(\eta^n)] - \theta[F^{n+1} \circ \mathcal{B}(\eta^{n+1})], \tag{2.9}$$

and set $\gamma^{n+1} = \mathcal{B}(\eta^{n+1})$.

REMARK 2.1. In the case of $\theta = 0$, the right-hand side of (2.9) does not involve η^{n+1} and, since M_h is diagonal, an explicit solution of the *discrete problem* is given by the time-stepping algorithm

$$\begin{aligned}\eta^{n+1} &= \eta^n - \Delta t_{n+1} M_h^{-1} F^n(\gamma^n),\\ \gamma^{n+1} &= \mathcal{B}(\eta^{n+1}).\end{aligned}$$

For $\theta > 0$, in contrast, (2.9) is a nonlinear system of equations, since the maps $\mathcal{B}$ and F^{n+1} are nonlinear.

3. Solving the nonlinear system. As the system (2.9) is nonlinear, no direct solution formula is available. In this section, we employ the theory of piecewise affine functions to prove the existence and uniqueness of a solution to this system and the convergence of an algorithm to find the solution.

3.1. Piecewise affine functions. Following Scholtes [16] we define the following.

DEFINITION 3.1. *A continuous function $f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ is called* piecewise affine *if there exists a finite set of functions $f_i(x) = A^i x + b^i$, $i = 1, \ldots, k$, such that $f(x) \in \{f_1(x), \ldots, f_k(x)\}$ for every $x \in \mathbb{R}^n$. The set of pairs (A^i, b^i), $i = 1, \ldots, k$, is called a collection of* matrix-vector pairs *corresponding to f.*

The following theorem of Fujisawa and Kuh [7] is the central tool of this section. In this paper, we formulate it using the terminology of Scholtes. Recall that a homeomorphism is a continuous bijective map with a continuous inverse.

THEOREM 3.2. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be piecewise affine, and let $(A^1, b^1), \ldots, (A^k, b^k)$ be a collection of matrix-vector pairs corresponding to f. We denote by A_l^i the submatrix of A^i composed of its first l columns and rows. If, for each $l = 1, \ldots, n$, the determinants of*

$$A_l^1, \ldots, A_l^k$$

do not vanish and have the same sign, then f is a homeomorphism.

To apply the theorem, we need the following result from Gantmacher and Krein [8] about tridiagonal matrices of the form

$$T = \begin{bmatrix} a_1 & b_1 & & & 0 \\ c_1 & a_2 & \ddots & & \\ & c_2 & \ddots & b_{n-2} & \\ & & \ddots & a_{n-1} & b_{n-1} \\ 0 & & & c_{n-1} & a_n \end{bmatrix}. \tag{3.1}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

THEOREM 3.3. *Let T be tridiagonal with the property*

$$b_i, c_i < 0 \quad \text{for all } i \in \{1, \dots, n-1\},$$

where b_i and c_i are defined as in (3.1). Then, all eigenvalues of T are real and simple.

In addition, the following simple consequence of the Gershgorin circle theorem will be employed.

COROLLARY 3.4. *Let $A \in \mathbb{R}^{n\times n}$ with $a_{ii} > 0$ for all $i \in \{1, \dots, n\}$ be strictly column-wise diagonally dominant. Then every eigenvalue λ of A satisfies $\Re(\lambda) > 0$.*

Together, Theorem 3.3 and Corollary 3.4 yield the following result.

PROPOSITION 3.5. *Let T be a tridiagonal matrix of the form (3.1) such that the conditions of Corollary 3.4 are satisfied and the off-diagonal entries in each column are either all zero or all negative. Then all leading principal minors of T are strictly positive.*

Proof. Let $\hat{T}$ be a submatrix of T composed of its first k columns and rows. We need to show that $\det(\hat{T}) > 0$. All assumptions imposed on T also hold for $\hat{T}$.

The first step of the proof is to reduce the problem to the case $c_i b_i < 0$ for all $i \in \{1, \dots, n\}$. Assume that there exists an $i \in \{1, \dots, n\}$ such that the off-diagonal entries of the ith column vanish. Then, $\hat{T}$ has the structure shown in Figure 3.1, where T_1 and T_2

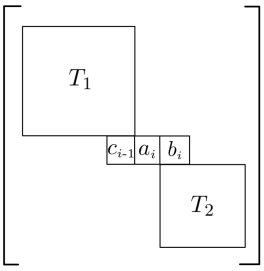


FIG. 3.1. *Structure of matrix $\hat{T}$.*

are square tridiagonal matrices that also satisfy the assumptions imposed on T. Recall that any eigenvalue of T is a complex root of $\rho_{\hat{T}}(\lambda) = \det(\hat{T} - \lambda I)$. By the above form of $\hat{T}$, expanding this determinant along column i using the Laplace expansion formula yields

$$\rho_{\hat{T}}(\lambda) = \det(T_1 - \lambda I)\det(T_2 - \lambda I)(a_{ii} - \lambda).$$

Hence, the spectral decomposition

$$\lambda(\hat{T}) = \lambda(T_1) \cup \lambda(T_2) \cup \{a_{ii}\}$$

holds. Now we apply the same decomposition to T_1 and T_2 if columns with vanishing off-diagonal entries exist in those matrices, and we continue recursively.

It remains to show the case $c_i, b_i < 0$ for all $i \in \{1, \dots, n\}$. Theorem 3.3 implies that all eigenvalues of $\hat{T}$ are real and Corollary 3.4 further establishes the strict positivity of these eigenvalues. Thus,

$$\det(\hat{T}) = \prod_{\lambda \in \lambda(\hat{T})} \lambda > 0,$$

completing the proof. ∎

ETNA
Kent State University and
Johann Radon Institute (RICAM)

3.2. Homeomorphism property of the implicit system. Consider the function

$$\begin{aligned}\Phi : \mathbb{R}^\kappa &\longrightarrow \mathbb{R}^\kappa \\ x &\longmapsto \frac{1}{\Delta t_{n+1}} M_h(x-\eta^n) + \theta F^{n+1} \circ \mathcal{B}(x) + (1-\theta)[F^n \circ \mathcal{B}(\eta^n)].\end{aligned} \tag{3.2}$$

Being a solution of (2.9) is equivalent to being a root of Φ. Thus, to apply the above theory we need to establish that

(i) the function Φ is piecewise affine, and
(ii) the corresponding matrices fulfill the conditions of Proposition 3.5.

Then, by Proposition 3.5, the assumptions of Theorem 3.2 are fulfilled, which proves that Φ is a homeomorphism.

The map Φ is continuous because the function $\beta(\eta, x_i)$ is continuous in η, and, for the term $k_i(\beta(\eta, x_i))\beta(\eta, x_i)$ in (2.6), the one-sided limits $\eta \downarrow L_i$ and $\eta \uparrow L_i$ at $\eta = L_i$ are both zero. To show the remaining part of (i) as well as (ii), we explore the nonlinear function $F^n \circ \mathcal{B}$. We partition $\mathbb{R}^\kappa$ into box-shaped regions as follows. Consider the map

$$\begin{aligned}q_i : \mathbb{R} &\longrightarrow \{-1,\, 0,\, 1\}, \\ \eta &\longmapsto \begin{cases} -1 & \text{for } \eta < 0, \\ 0 & \text{for } 0 \le \eta \le L_i, \\ 1 & \text{for } L_i < \eta, \end{cases}\end{aligned} \tag{3.3}$$

with $i \in \{1, \ldots, \kappa\}$. In addition, define the set

$$\mathcal{Z} := \{-1, 0, 1\}^\kappa, \tag{3.4}$$

and, for each element $(z_1, \ldots, z_\kappa) \in \mathcal{Z}$, the convex region

$$P_z := \bigotimes_{i=1}^{\kappa} q_i^{-1}(z_i) \subseteq \mathbb{R}^\kappa. \tag{3.5}$$

For example, $P_{(1,0,1)} = (L_1, \infty) \times [0, L_2] \times (L_3, \infty)$ (see also Figure 3.2). It can be easily checked that $(P_z)_{z \in \mathcal{Z}}$ is a partition of $\mathbb{R}^\kappa$.

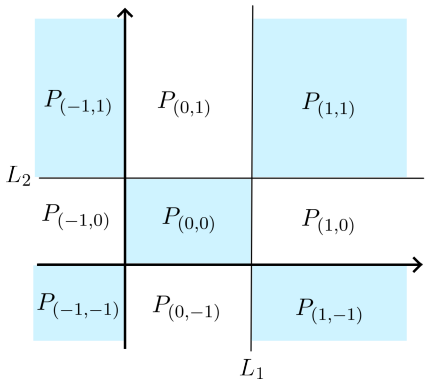


FIG. 3.2. *Illustration of the partitioning of* $\mathbb{R}^2$ *by* (3.3), (3.4), *and* (3.5).

To simplify notation in the matrix definition (3.6) in the theorem below, instead of placing z_i as a superscript on each variable of the ith column for all $i \in \{1, \ldots, \kappa\}$, the common

ETNA
Kent State University and
Johann Radon Institute (RICAM)

superscript is written above each column. In addition, $\mathbf{1}_{x=y}$ denotes the indicator function $\mathbf{1}_A(x)$ on the set $A := \{y\}$.

THEOREM 3.6. *Let $z \in \mathcal{Z}$. For all $\eta \in P_z$ it holds that*

$$F^n \circ \mathcal{B}(\eta) = A(z)[\eta - b(z)] - c^n,$$

where

$$A(z) := \begin{bmatrix} \overset{z_1}{[r_1+r_2]g_1} & \overset{z_2}{-r_2 g_2} & & \overset{z_{\kappa-1}}{} & \overset{z_\kappa}{0} \\ -r_2 g_1 & [r_2+r_3]g_2 & \ddots & & \\ & -r_3 g_2 & \ddots & -r_{\kappa-1}g_{\kappa-1} & \\ & & \ddots & [r_{\kappa-1}+r_\kappa]g_{\kappa-1} & -r_\kappa g_\kappa \\ 0 & & & -r_\kappa g_{\kappa-1} & r_\kappa g_\kappa \end{bmatrix}, \tag{3.6}$$

with

$$r_i^z := \frac{k_i(z)}{h_i}, \quad g_i^z := \frac{\mathbf{1}_{z=-1}}{c_i^f} + \frac{\mathbf{1}_{z=1}}{c_i^u}, \quad \text{for } i = 1, \dots, \kappa,\ z \in \{-1, 0, 1\},$$

$$b(z) := [\mathbf{1}_{z_1=1}L_1,\ \dots,\ \mathbf{1}_{z_\kappa=1}L_\kappa]^{\mathrm{T}},$$

and

$$c^n := [r_1^{\operatorname{sgn}(s^n)} s^n,\ 0,\ \dots,\ 0]^{\mathrm{T}}.$$

Proof. Let $z \in \mathcal{Z}$ and $\eta \in P_z$. Additionally, define $\gamma := \mathcal{B}(\eta)$. We begin by computing a matrix form of $\mathcal{B}$. Let $i \in \{1, \dots, \kappa\}$. Since $\eta \in P_z$, we have $q_i(\eta_i) = z_i$ and thus it holds that

$$z_i = 1 \iff \eta_i > L_i \quad \text{and} \quad z_i = -1 \iff \eta_i < 0.$$

Recalling (2.2) we obtain

$$\begin{aligned} \mathcal{B}(\eta)_i = \beta(\eta_i, x_i) &= \mathbf{1}_{\eta_i<0}\frac{\eta_i}{c_i^f} + \mathbf{1}_{\eta_i>L_i}\frac{\eta_i - L_i}{c_i^u} \\ &= \mathbf{1}_{z_i=-1}\frac{\eta_i}{c_i^f} + \mathbf{1}_{z_i=1}\frac{\eta_i - L_i}{c_i^u} \\ &= \mathbf{1}_{z_i=-1}\frac{\eta_i - \mathbf{1}_{z_i=1}L_i}{c_i^f} + \mathbf{1}_{z_i=1}\frac{\eta_i - \mathbf{1}_{z_i=1}L_i}{c_i^u} \\ &= \left[\frac{\mathbf{1}_{z_i=-1}}{c_i^f} + \frac{\mathbf{1}_{z_i=1}}{c_i^u}\right][\eta_i - \mathbf{1}_{z_i=1}L_i] = g_i^{z_i}[\eta_i - b(z)_i]. \end{aligned}$$

It follows that

$$\mathcal{B}(\eta) = \operatorname{diag}(g_1^{z_1}, \dots, g_\kappa^{z_\kappa})[\eta - b(z)].$$

To obtain a matrix form of F^n, we recall (2.8) and (2.7) and calculate

$$F^n(\gamma) = \begin{bmatrix} Q_1^n(\gamma) - Q_2^n(\gamma) \\ \vdots \\ Q_{\kappa-1}^n(\gamma) - Q_\kappa^n(\gamma) \\ Q_\kappa^n(\gamma) \end{bmatrix} = \overbrace{\begin{bmatrix} 1 & -1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & -1 \\ 0 & & & 1 \end{bmatrix}}^{:=\mathbb{I}} \begin{bmatrix} Q_1^n(\gamma) \\ \vdots \\ Q_\kappa^n(\gamma) \end{bmatrix}.$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Recalling (2.6) and noting that $k_i(\gamma_j)/h_i = r_i^{z_j}$, the expression reads

$$\begin{aligned}
&\mathbb{I} \times \begin{bmatrix} [k_1(\gamma_1)\gamma_1 - k_1(s^n)s^n]/h_1 \\ [k_2(\gamma_2)\gamma_2 - k_2(\gamma_1)\gamma_1]/h_2 \\ \vdots \\ [k_\kappa(\gamma_\kappa)\gamma_\kappa - k_\kappa(\gamma_{\kappa-1})\gamma_{\kappa-1}]/h_\kappa \end{bmatrix} \\
&= \mathbb{I} \times \begin{bmatrix} r_1^{z_1} & & & 0 \\ -r_2^{z_1} & r_2^{z_2} & & \\ & \ddots & \ddots & \\ 0 & & -r_\kappa^{z_{\kappa-1}} & r_\kappa^{z_\kappa} \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_\kappa \end{bmatrix} - c^n \\
&\qquad\qquad z^1 \qquad z^2 \qquad \cdots \qquad z^\kappa \\
&= \begin{bmatrix} r_1 + r_2 & -r_2 & & 0 \\ -r_2 & r_2 + r_3 & \ddots & \\ & \ddots & \ddots & -r_\kappa \\ 0 & & -r_\kappa & r_\kappa \end{bmatrix} \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_\kappa \end{bmatrix} - c^n.
\end{aligned}$$

The result follows from substituting $\gamma = \mathcal{B}(\eta)$. ∎

Theorem 3.6 implies that Φ is piecewise affine with corresponding matrices

$$J_z := \frac{1}{\Delta t_{n+1}} M_h + \theta A(z),$$

and thus property (i) is established.

To conclude the proof of this section, it remains to show property (ii). The off-diagonal entries of the ith column of J_z are $-\theta r_i^{z_i} g_i^{z_i}$, $-\theta r_{i+1}^{z_i} g_i^{z_i}$ or both. Hence, if $z_i \in \{-1, 1\}$, these values are jointly negative by (2.5). On the other hand, if $z_i = 0$, these entries are jointly zero. In addition, the diagonal of J_z is

$$(J_z)_{ii} = (M_h)_{ii}/\Delta t_{n+1} + \theta r_i^{z_i} g_i^{z_i} + \theta r_{i+1}^{z_i} g_i^{z_i}$$

for $i = 1, \ldots, \kappa - 1$ and

$$(J_z)_{ii} = (M_h)_{ii}/\Delta t_{n+1} + \theta r_i^{z_i} g_i^{z_i}$$

for $i = \kappa$. These numbers are strictly positive and strictly larger than the sum of the moduli of the off-diagonal entries. Thus, we have proven property (ii), and the main result of this section follows.

THEOREM 3.7. *The map Φ as defined in (3.2) is a homeomorphism. In particular, (2.9) has a unique solution η^{n+1}.*

3.3. Katzenelson algorithm for the implicit system. We apply the Katzenelson algorithm [9] to find a root of Φ. Fujisawa and Kuh [7] prove that this algorithm finds the exact root of a function in a finite number of steps from any initial guess, provided that the function is a piecewise affine homeomorphism and an appropriate corner scheme is implemented. For the reader's convenience, we adapt the method to the setting of this study and provide a self-contained convergence proof under simplifying assumptions. Under these assumptions, the discussion of corner schemes can be avoided while convergence from almost every initial

ETNA
Kent State University and
Johann Radon Institute (RICAM)

guess is retained. The proof follows the ideas presented in [7]. Example iterates are shown in Figure 3.3.

Algorithm 1 Katzenelson algorithm.

Initialization. Let $x^{(0)} \in P_{z^{(0)}}$ be an initial guess, where $z^{(0)} \in \mathcal{Z}$. We refer to $P_{z^{(0)}}$ as the *active region*. Set $k := 0$.

Stopping criterion. If $\Phi(x^{(k)}) = 0$, stop the algorithm. Otherwise, continue with the next iteration.

Iteration step. Compute the Newton direction $v^{(k)} \in \mathbb{R}^\kappa$ by solving

$$J_{z^{(k)}} v^{(k)} = -\Phi(x^{(k)}). \tag{3.7}$$

Determine the maximal step length for which the iterate remains in the closure of the active region:

$$\lambda^{(k)} := \max\{\lambda \in [0,1] \mid x^{(k)} + \lambda v^{(k)} \in \overline{P_{z^{(k)}}}\}, \tag{3.8}$$

and set

$$x^{(k+1)} := x^{(k)} + \lambda^{(k)} v^{(k)}.$$

If $\lambda^{(k)} = 1$, then the full Newton step remains in $\overline{P_{z^{(k)}}}$. Consequently,

$$\Phi(x^{(k+1)}) = 0,$$

as shown by (3.11). Stop the algorithm.

Suppose now that $\lambda^{(k)} < 1$. Then, $x^{(k+1)}$ lies on the boundary of the active region. Thus, by construction of the box partition, the point belongs to the closure of at least two regions. We now assume that $x^{(k+1)}$ belongs to the closure of no more than two regions. Then, there is a unique index $z \in \mathcal{Z} \setminus \{z^{(k)}\}$ such that

$$x^{(k+1)} \in \overline{P_z}.$$

Set

$$z^{(k+1)} := z,$$

which makes $P_{z^{(k+1)}}$ the new active region. Set $k := k+1$ and return to the stopping criterion.

Following the terminology of [7], we call a point that lies in the closure of more than two regions a corner. We remark that the no-corner assumption made in the algorithm is not needed for the algorithm to function if the corner scheme developed in [7] is implemented. For simplicity, we do not describe or analyze this implementation here.

The analysis of the algorithm relies on considering the line segment

$$L := \overline{\Phi(x^{(0)})\, 0}$$

in the image space $\Phi(\mathbb{R}^\kappa) = \mathbb{R}^\kappa$ that connects $\Phi(x^{(0)}) \in \mathbb{R}^\kappa$ and $0 \in \mathbb{R}^\kappa$. Its preimage $\Phi^{-1}(L)$ is called the *solution curve*. As we will see, the algorithm traces this curve exactly.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

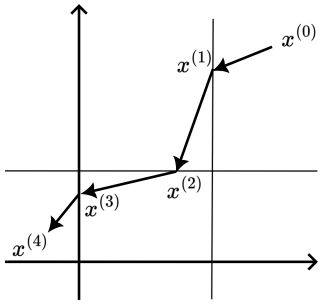


FIG. 3.3. *Example iterates of the Katzenelson algorithm. Compare with Figure 3.2.*

We make the following assumptions:

1. The initial guess $x^{(0)}$ belongs to the interior of a region $P_{z^{(0)}}$.
2. Every point of $L \setminus \{0\}$ belongs to the closure of at most two regions of the image partition $(\Phi(P_z))_{z\in\mathcal{Z}}$. That is,

$$N_y := \#\left\{z \in \mathcal{Z} \;\middle|\; y \in \overline{\Phi(P_z)}\right\} \leq 2 \quad \text{for every } y \in L \setminus \{0\}.$$

REMARK 3.8. Assumption 2 is essentially a reformulation of the no-corner assumption in the image space. However, the reformulation makes it clear that the assumption is not too restrictive in practice. Due to (2.5), the hyperplanes defining the box-shaped partition are distinct and therefore intersect in sets of codimension two. Taking the image of these intersections under Φ we observe that the set

$$E := \{y \in \mathbb{R}^\kappa \mid N_y > 2\}$$

is a union of sets of codimension two. In order for the line segment $\overline{y0} \setminus \{0\}$ to intersect E, the point y needs to be in the cone

$$C := \{\lambda w \mid \lambda \geq 0, w \in E\}.$$

Since the codimension of C is at least one, it has Lebesgue measure zero. Because Φ^{-1} is Lipschitz continuous, the set $\Phi^{-1}(C)$ also has Lebesgue measure zero. Therefore, Assumption 2 is satisfied for almost every initial guess $x^{(0)}$.

If the no-corner assumption is violated during a run, the algorithm may be restarted from a randomly perturbed point $\tilde{x}$ sampled from a probability measure that is absolutely continuous with respect to the Lebesgue measure. Then the new line segment

$$\overline{\Phi(\tilde{x})\,0} \setminus \{0\}$$

satisfies Assumption 2 almost surely. This strategy was implemented for the experiments in this study. Likewise, it is clear that Assumption 1 is satisfied for almost every $x^{(0)} \in \mathbb{R}^\kappa$. If this assumption is violated, a perturbation $\tilde{x}^{(0)}$ of the initial guess will again almost surely be admissible.

THEOREM 3.9. *Let Φ be defined by (3.2), and let Assumptions 1 and 2 hold for the initial guess $x^{(0)}$. Then Algorithm 1 finds a root of Φ in a finite number of steps.*

Proof. The proof is broken down into four steps.

Step 1: Well-definedness and solution-curve membership. We first show inductively that, for all $i = 0, 1, 2, \ldots$, the iterates $x^{(i)}$ of the algorithm (see Figure 3.3) are well-defined and

ETNA
Kent State University and
Johann Radon Institute (RICAM)

satisfy

$$\Phi(x^{(i)}) \in L. \tag{3.9}$$

For $i = 0$ the statement follows from the definition of L. Assume that $x^{(k)}$ is well-defined and $\Phi(x^{(k)}) \in L$ for some $k \in \mathbb{N}_0$. If $\Phi(x^{(k)}) = 0$, the stopping criterion terminates the algorithm. If

$$\Phi(x^{(k)}) \neq 0, \tag{3.10}$$

the algorithm proceeds as follows. The system (3.7) is solvable because $\det(J_{z^{(k)}}) > 0$ by the proof of Theorem 3.7. The maximum in (3.8) exists because it is taken over a nonempty, compact subset of $\mathbb{R}$. Indeed,

$$0 \in \left\{\lambda \in [0,1] \;\middle|\; x^{(k)} + \lambda v^{(k)} \in \overline{P_{z^{(k)}}}\right\} = [0,1] \cap g^{-1}\left(\overline{P_{z^{(k)}}}\right),$$

where

$$g(\lambda) := x^{(k)} + \lambda v^{(k)}$$

is continuous.

By the continuity of Φ, the affine formula for Φ on $P_{z^{(k)}}$ extends to $\overline{P_{z^{(k)}}}$. Since $x^{(k)} + \lambda^{(k)} v^{(k)} \in \overline{P_{z^{(k)}}}$ and $x^{(k)} \in \overline{P_{z^{(k)}}}$, we obtain

$$\begin{aligned} \Phi(x^{(k+1)}) &= \Phi(x^{(k)} + \lambda^{(k)} v^{(k)}) \\ &= \Phi(x^{(k)}) + \lambda^{(k)} J_{z^{(k)}} v^{(k)} \\ &= (1 - \lambda^{(k)}) \Phi(x^{(k)}). \end{aligned} \tag{3.11}$$

Therefore, $\Phi(x^{(k+1)}) \in L$. Additionally, if $\lambda^{(k)} = 1$, then the equality shows that $\Phi(x^{(k+1)}) = 0$, and the algorithm stops.

If $\lambda^{(k)} < 1$, the continuation of the algorithm requires that $x^{(k+1)}$ belong to the closure of exactly one region other than $P_{z^{(k)}}$. By the maximality of $\lambda^{(k)}$, we have

$$x^{(k+1)} \in \partial P_{z^{(k)}},$$

and, by the box structure of the partition, every boundary point of a region belongs to the boundary of a different region. Now consider the image regions

$$Q_z := \Phi(P_z)$$

given for all $z \in \mathcal{Z}$. If $z \in \mathcal{Z}$ satisfies $x^{(k+1)} \in \overline{P_z}$, then

$$\Phi(x^{(k+1)}) \in \Phi(\overline{P_z}) = \overline{Q_z},$$

because Φ is a homeomorphism. Since $\Phi(x^{(k+1)}) \in L$ and, by (3.11) and (3.10), also $\Phi(x^{(k+1)}) \neq 0$, Assumption 2 implies that $x^{(k+1)}$ belongs to the closures of at most two regions. Note that at this stage of the proof $\lambda^{(k)} = 0$ is still allowed, which could result in endless repeated switching between two regions.

Step 2: The region-exclusion property. Suppose now that the algorithm does not terminate. Then it produces $x^{(i)}$, $z^{(i)}$, $v^{(i)}$, and $\lambda^{(i)}$ for all $i \in \mathbb{N}_0$. By [16, Proposition 2.3.1], the inverse Φ^{-1} is also piecewise affine. Furthermore, the corresponding regions $(Q_z)_{z \in \mathcal{Z}}$ again partition $\mathbb{R}^\kappa$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Since the algorithm did not terminate immediately, we have

$$p := \Phi(x^{(0)}) \neq 0.$$

Therefore, we can parameterize L by the injective and continuous map

$$l : [0,1] \longrightarrow L \subseteq \mathbb{R}^\kappa, \quad s \longmapsto sp.$$

Since $\Phi(x^{(i)}) \in L$ by (3.9), we may define

$$s^{(i)} := l^{-1}(\Phi(x^{(i)}))$$

for all $i \in \mathbb{N}_0$.

Let $i \in \mathbb{N}_0$ and $s \in [0,1]$ such that $l(s) \in \overline{Q_{z^{(i)}}}$. Since the algorithm has not terminated at $x^{(i)}$, we have

$$0 \neq \Phi(x^{(i)}) = s^{(i)}p,$$

and hence $s^{(i)} \neq 0$. By the continuity of Φ^{-1}, the affine formula for Φ^{-1} on $Q_{z^{(i)}}$ extends to $\overline{Q_{z^{(i)}}}$. Therefore, we have

$$\begin{aligned}\Phi^{-1}(l(s)) - x^{(i)} &= \Phi^{-1}(sp) - \Phi^{-1}(s^{(i)}p) \\ &= J_{z^{(i)}}^{-1}(s - s^{(i)})p = \frac{s^{(i)} - s}{s^{(i)}} v^{(i)}.\end{aligned}$$

We note that

$$\Phi^{-1}(l(s)) \in \Phi^{-1}\big(\overline{Q_{z^{(i)}}}\big) = \overline{P_{z^{(i)}}}.$$

In addition, if $s \leq s^{(i)}$, then $(s^{(i)} - s)/s^{(i)} \in [0,1]$. Thus, in that case, the quotient is admissible in the maximization (3.8), and the maximality of $\lambda^{(i)}$ implies that

$$\lambda^{(i)} \geq \frac{s^{(i)} - s}{s^{(i)}}. \tag{3.12}$$

If $s > s^{(i)}$, the inequality (3.12) holds trivially because $\lambda^{(i)} \geq 0$ by construction.

Furthermore, (3.11) implies that

$$s^{(i+1)}p = (1 - \lambda^{(i)})s^{(i)}p,$$

and therefore

$$\lambda^{(i)} = \frac{s^{(i)} - s^{(i+1)}}{s^{(i)}}, \tag{3.13}$$

because $p \neq 0$. Combining this equality with (3.12), we obtain

$$s \geq s^{(i+1)}. \tag{3.14}$$

We have thus shown that, for every $s \in [0,1]$,

$$l(s) \in \overline{Q_{z^{(i)}}} \quad \Longrightarrow \quad s \geq s^{(i+1)},$$

or, equivalently,

$$s < s^{(i+1)} \quad \Longrightarrow \quad l(s) \notin \overline{Q_{z^{(i)}}}. \tag{3.15}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Step 3: Strict decrease of the path parameter. We now show that

$$s^{(0)} > s^{(1)} > s^{(2)} > \cdots . \tag{3.16}$$

To show $s^{(0)} > s^{(1)}$, we note that, by Assumption 1 and the continuity of Φ^{-1},

$$l(1) = \Phi(x^{(0)}) \in \operatorname{int}(Q_{z^{(0)}}).$$

Hence, we may choose $s < 1$ sufficiently close to 1 such that

$$l(s) \in Q_{z^{(0)}}.$$

Using (3.14), we obtain

$$s^{(1)} \leq s < 1 = s^{(0)}.$$

Now let $k \geq 1$. Define

$$y := \Phi(x^{(k)}) = l(s^{(k)}).$$

By construction,

$$y \in \overline{Q_{z^{(k-1)}}} \cap \overline{Q_{z^{(k)}}}.$$

Since $y \neq 0$, Assumption 2 implies that these are the only region closures containing y. Since, additionally, $\mathcal{Z}$ is finite and all sets $\overline{Q_z}$ are closed, we can set

$$\tilde{Z} := \left\{ z \in \mathcal{Z} \;\middle|\; z \neq z^{(k-1)},\ z \neq z^{(k)} \right\}$$

and obtain

$$\varepsilon := \tfrac{1}{2} \min\left\{ \operatorname{dist}\left(\{y\}, \overline{Q_z}\right) \;\middle|\; z \in \tilde{Z} \right\} > 0.$$

The ball $B_\varepsilon(y)$ then lies in the complement of $\bigcup_{z \in \tilde{Z}} \overline{Q_z}$, and hence

$$B_\varepsilon(y) \subseteq \overline{Q_{z^{(k-1)}}} \cup \overline{Q_{z^{(k)}}}.$$

Choose $s < s^{(k)}$ sufficiently close to $s^{(k)}$ such that

$$l(s) \in B_\varepsilon(y).$$

By (3.15),

$$l(s) \notin \overline{Q_{z^{(k-1)}}},$$

and hence

$$l(s) \in \overline{Q_{z^{(k)}}}.$$

Applying (3.14) again, we obtain

$$s^{(k+1)} \leq s < s^{(k)}.$$

Since $k \geq 1$ was arbitrary, we have shown (3.16). We note that (3.16) together with (3.13) imply that $\lambda^{(k)} > 0$ for all $k \in \mathbb{N}_0$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Step 4: Finite termination. Let $k \geq 1$. Using (3.16) for any $j \leq k-2$, we obtain $s^{(k)} < s^{(j+1)}$. Thus, by (3.15),

$$\Phi(x^{(k)}) = l(s^{(k)}) \notin \overline{Q_{z^{(j)}}}.$$

On the other hand,

$$\Phi(x^{(k)}) \in \overline{Q_{z^{(k-1)}}}, \quad \Phi(x^{(k)}) \in \overline{Q_{z^{(k)}}},$$

and $z^{(k-1)} \neq z^{(k)}$ by construction. Hence, $z^{(k)}$ is distinct from $z^{(0)}$, $z^{(1)}$, $\ldots$, $z^{(k-1)}$. Since $k \geq 1$ was arbitrary, all

$$z^{(0)}, z^{(1)}, z^{(2)}, \ldots$$

are pairwise distinct. This contradicts the finiteness of $\mathcal{Z}$, and completes the proof. ∎

REMARK 3.10. Due to floating-point errors resulting from each step of the algorithm, especially from solving (3.7), we cannot expect the stopping criterion to be satisfied exactly in practice. Therefore, we suggest replacing the criterion with the tolerance-based condition

$$\|\Phi(x^{(k)})\| \leq \mathrm{t}_r \|\Phi(x^{(0)})\| + \mathrm{t}_a,$$

where t_r and t_a are predefined relative and absolute tolerances, and $\|\cdot\|$ is the Euclidean norm. This criterion was also used in the implementations of the numerical experiments shown below. See Appendix B for pseudocode for the implementation.

4. Results. In this section, we present the results of numerical experiments conducted with the proposed method and compare them with the Neumann analytical solution of the Stefan problem.

4.1. Neumann analytical solution. The Neumann analytical solution [17] is valid for the Stefan problem on a semi-infinite, homogeneous, one-dimensional domain with a Dirichlet boundary condition. Initially, the material is unfrozen at the uniform temperature $T_i > T_f$, while the boundary temperature is fixed at $T_0 < T_f$ for $t > 0$. Here, T_f denotes the phase-change temperature. The solution [17] is then given by

$$u(x,t) = \begin{cases} T_0 + (T_f - T_0)\dfrac{\operatorname{erf}\big(x/(2\sqrt{\alpha_f t})\big)}{\operatorname{erf}\big(\xi\sqrt{b}\big)}, & 0 \leq x \leq X(t), \\ T_i - (T_i - T_f)\dfrac{\operatorname{erfc}\big(x/(2\sqrt{\alpha_u t})\big)}{\operatorname{erfc}(\xi)}, & x > X(t), \end{cases}$$

where

$$X(t) = 2\xi\sqrt{\alpha_u t},$$

and

$$\alpha_f := k_f/c_f, \quad \alpha_u := k_u/c_u, \quad b := \alpha_u/\alpha_f.$$

The dimensionless constant $\xi > 0$ is the unique positive solution of

$$b_4 F_2\big(\sqrt{b}\,\xi\big) - b_3 F_1(\xi) - \xi = 0,$$

with

$$F_1(y) := \frac{\exp(-y^2)}{\operatorname{erfc}(y)}, \qquad F_2(y) := \frac{\exp(-y^2)}{\operatorname{erf}(y)},$$
$$b_3 := \frac{c_u(T_i - T_f)}{L\sqrt{\pi}}, \qquad b_4 := \frac{k_f(T_f - T_0)}{L\sqrt{\pi\alpha_f\alpha_u}}.$$

For all comparisons we used $T_f = 0$.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

4.2. Numerical experiments. Numerical benchmarking experiments were performed using the backward Euler method ($\theta = 1$), the Crank–Nicolson scheme ($\theta = 1/2$), and the forward Euler method ($\theta = 0$). The physical parameters used were

$$k_f = 1.233783\ \mathrm{W\,m^{-1}\,K^{-1}}, \qquad k_u = 0.730975\ \mathrm{W\,m^{-1}\,K^{-1}},$$
$$c_f = 9.1018 \times 10^5\ \mathrm{J\,m^{-3}\,K^{-1}}, \quad c_u = 1.43518 \times 10^6\ \mathrm{J\,m^{-3}\,K^{-1}},$$
$$L = 7.5 \times 10^7\ \mathrm{J\,m^{-3}}.$$

The prescribed boundary and initial temperatures were $T_0 = -25\,^\circ\mathrm{C}$ and $T_i = 10\,^\circ\mathrm{C}$, respectively. Since a semi-infinite domain cannot be simulated, the domain was truncated at $5\,\mathrm{m}$ depth and a zero-flux Neumann condition was used at $5\,\mathrm{m}$.

For all three variants of the scheme, the computed results closely matched the Neumann analytical solution, provided that a sufficiently refined spatial mesh and a suitably small time step were used. The results of the implicit method for increasing spatial resolution are presented in Figure 4.1. Additionally, Figure 4.2 depicts the convergence of the numerical error towards zero as the mesh is refined.

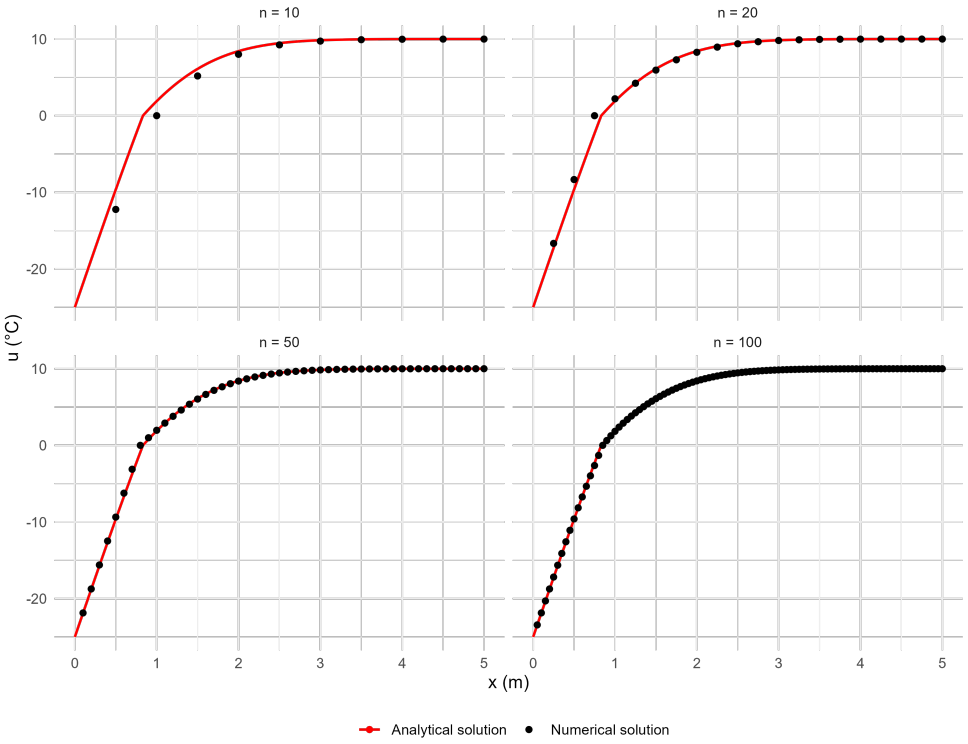


FIG. 4.1. *Convergence of numerical results to the analytical temperature distribution on day 15 with increasing spatial resolution. The temporal resolution is fixed at 10,000 time steps per day. Temperature is denoted by u.*

The backward Euler method ($\theta = 1$) and the Crank–Nicolson scheme ($\theta = 0.5$), in combination with the Katzenelson solver, were also implemented in LPJmL 5 [24] to assess their robustness and stability. A 13 m-deep soil column was discretized for each grid cell (more than 60,000 in total) using an unequally spaced grid that becomes progressively coarser towards the bottom ($n = 24$). The model uses a daily time step. The GSWP3-W5E5 dataset [10]

ETNA
Kent State University and
Johann Radon Institute (RICAM)

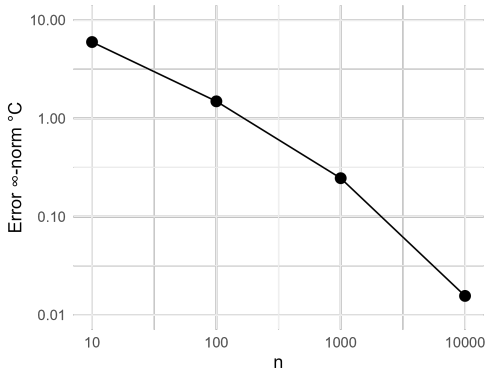


FIG. 4.2. *Discretization error in the temperature distribution for all 20 days with increasing resolution. The error was measured using the l^∞ norm over time and space.*

was used as climate forcing. The lower mean air temperature of these data served as the top Dirichlet boundary condition. The thermal influence of snow cover within a grid cell was accounted for in a simplified manner by dynamically increasing the thickness and decreasing the thermal conductivity of the top layer in the presence of snow. To optimize computational efficiency, the Crank–Nicolson scheme for the standard heat equation was applied whenever all temperatures in the soil column as well as the boundary condition had the same sign.

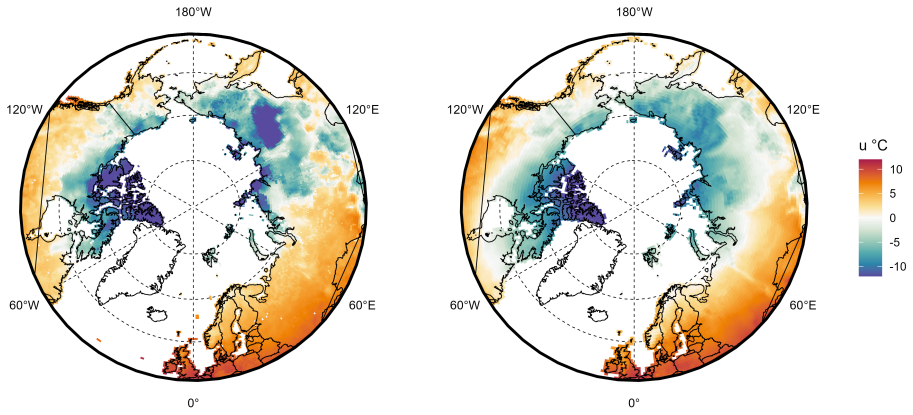


FIG. 4.3. *Mean soil temperature υ at the depth of zero annual amplitude—the shallowest depth at which annual ground-temperature fluctuations are practically zero—for 2000 to 2014 in LPJmL (left), compared with statistically interpolated observations [1] (right).*

The Katzenelson method successfully converged for all grid cells throughout the entire simulation period of over 4,000 years for both the backward Euler and Crank–Nicolson schemes. Figure 4.3 presents the resulting deep soil temperatures obtained using the backward Euler method, in comparison with observational data. Since multiple iterations may be required to find the root of Φ, the method is computationally more expensive than solving a linear heat equation. However, in the global LPJmL simulation, a single iteration was sufficient in most cases, indicating that no node changed state during the time step (see Figure 4.4 for the backward Euler method). During the first transient simulation year, an average of 1.48

ETNA
Kent State University and
Johann Radon Institute (RICAM)

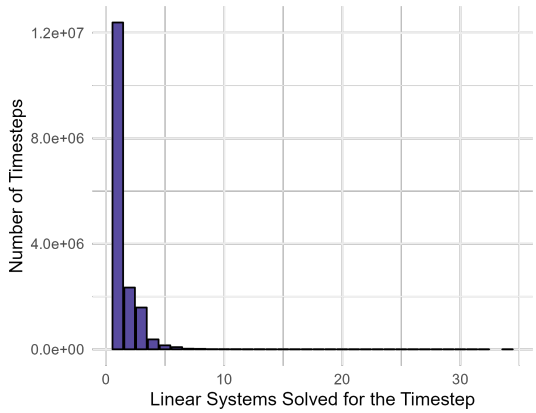


FIG. 4.4. *Number of time steps (y-axis) that needed* $1, 2, 3, \ldots, 35$ *linear system solutions (x-axis) to carry out the step. The first year of the LPJmL simulation* ($n = 24$) *using the backward Euler method is shown.*

and 1.93 linear system solutions per time step were required for the backward Euler and Crank–Nicolson schemes, respectively. Experiments with increasing spatial resolution in LPJmL as well as in the benchmark problem suggest that the iteration count per time step grows linearly for the benchmark problem and slightly faster than linearly in the LPJmL setup (see Figure 4.5).

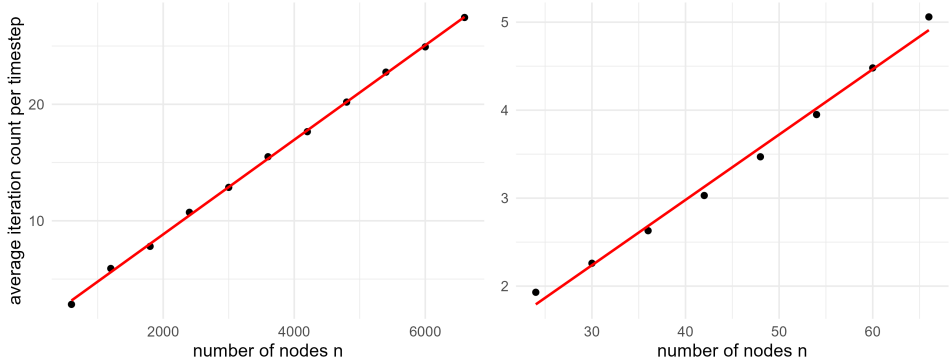


FIG. 4.5. *Average number of iterations (linear system solutions) needed per time step for different spatial resolutions. The left-hand side shows the results for the benchmark problem used to compare with analytical solutions. The right-hand side shows the averaged results for a 100-year LPJmL simulation with climate data. Both experiments used* $\theta = 0.5$. *The fitted linear regression line is shown for reference in red.*

5. Discussion. The observed dependence of the iteration count on the number of nodes deserves discussion. The proposed scheme is an instance of the broader class of piecewise linear homotopy methods discussed by Todd [18]. The homotopy corresponding to the scheme is

$$h(x,t) := t\Phi(x) + (1-t)\Phi_0(x),$$

where

$$\Phi_0(x) := \Phi(x) - \Phi(x^{(0)}).$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

For such algorithms, Todd showed that the number of solver iterations can be as large as 2^{n-1}, even when both Φ and Φ_0 are linear [19]. However, Todd acknowledges that the Katzenelson method is immune to the type of construction he used because those constructions require $J\Phi(x) \neq J\Phi_0(x)$. In our experiments, we indeed observed a much slower growth of the iteration count than 2^{n-1} (see Figure 4.5). Nevertheless, these observations do not rule out exponential complexity as the worst-case behavior when using, for example, different input data, a much larger time step, or a different grid.

In comparison with the Gauss–Seidel-type method of White [27], the numerical experiments with our implementation of the Katzenelson method suggest that fewer iterations per time step could be required to reach convergence. In contrast to the 12–18 iterations for $n = 21$ reported by White, the average iteration count in LPJmL using this implementation was only 1.48 for $n = 24$. However, these results may not be comparable due to the use of a different time step and different input data. In addition, each iteration of White's method requires less computational effort than the method proposed here, as no linear system needs to be solved.

We also remark that, up to floating-point error, our solver finds the exact discrete solution to the system rather than a numerical approximation to this solution. If no node changes state during an iteration (i.e., the freezing front does not move), the initial guess already lies in the same region as the solution. Thus, in those cases, the exact solution is found in a single iteration. Finally, in contrast to White's method, our method does not impose a time-step restriction for its convergence guarantees. It is worth emphasizing, however, that all these improvements come at the cost of allowing only piecewise constant thermal conductivity and a piecewise affine temperature-enthalpy relation rather than fully nonlinear functions.

As mentioned in the introduction, the most commonly used implicit DECP method for soil heat transfer in LSMs and DGVMs neglects the nonlinearity in the partial differential equation during the heat diffusion time step. It only corrects for the effects of phase changes after the diffusion time step has been fully carried out. This approach has several implications when compared to the method proposed in this study.

First, the DECP method requires solving only a single linear system per time step, making it computationally more efficient than methods that handle a nonlinear system for time stepping. However, by leveraging the piecewise affine structure of the system, the method proposed in this study usually requires solving only about 1.5 to 2 linear systems per time step with a typical choice of n, which remains computationally feasible for global models. While the backward Euler method required fewer iterations per time step than the Crank–Nicolson scheme, the latter's higher temporal order of accuracy may well justify its use.

Second, while the implicit DECP method is computationally more efficient, it is known to introduce an artificial stretching of the freezing region, leading to significant errors [13]. This effect was reproduced in the present study (see Figure 5.1). Over the full 20-day simulation period, the mean absolute difference between the analytical and numerical solutions was $0.443\,^{\circ}\mathrm{C}$ for the DECP method, whereas the proposed method achieved a lower error of $0.170\,^{\circ}\mathrm{C}$. Furthermore, it was observed that the explicit DECP scheme does not exhibit this nonphysical behavior and that the discrepancy increases with larger time steps.

The errors in the implicit DECP method can be attributed to the fact that, in implicit schemes, energy fluxes are computed based on the temperature distribution at the next time step. Since the phase-change parameterization is applied separately, the predicted temperature distribution is not yet adjusted for phase changes, leading to inaccuracies. This explanation is supported by the observation that the explicit DECP scheme does not exhibit this issue, as the fluxes in this approach are derived from the current temperature distribution, which has already been adjusted for phase changes.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

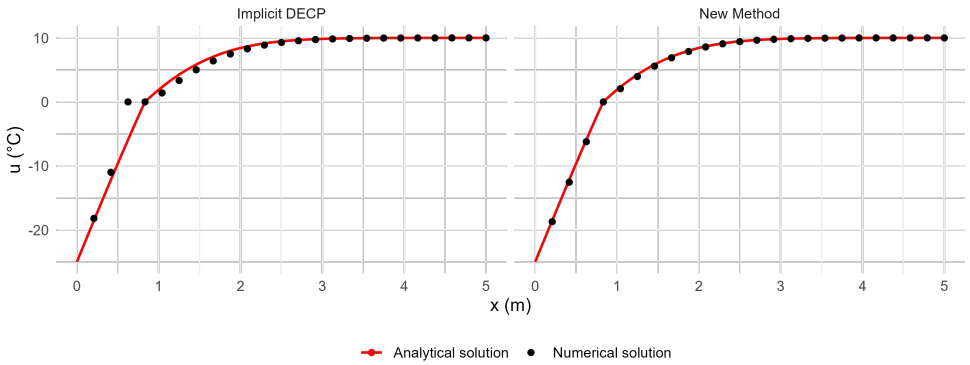


FIG. 5.1. *Implicit DECP method and new method proposed here plotted against the analytical solution on day 15. Temperature is denoted by u. Both methods used $\theta = 1/2$. The time step was set to one day.*

A potential approach to enhance the computational performance of the implicit scheme developed in this study is to use a predictor-corrector method. In this approach, an explicit time step could be used to generate the initial guess for the Katzenelson algorithm. It is plausible that this initial guess would, in most cases, already lie within the same region as the solution, thereby reducing the number of iterations needed to find the solution to one in those cases.

6. Conclusion. We proposed an implicit solver for one-dimensional heat conduction with phase change in global DGVMs and LSMs. The Katzenelson algorithm finds the exact solution of the nonlinear time-stepping system in a finite number of iterations for almost every initial guess, with no restriction on the time step. In contrast, the convergence guarantee of White's method [27] depends on a time-step bound that is not explicitly known. Our guarantee requires piecewise constant thermal conductivities and a piecewise affine temperature-enthalpy relation, as is standard in permafrost modeling.

Tests in LPJmL confirmed this robustness in practice. At typical model resolutions, the computational cost remained within a factor of two of that of the implicit DECP method. Unlike DECP, the new method does not decouple heat conduction from phase change, and therefore avoids the artificial extension of the freezing region. In summary, it is more accurate than DECP while remaining computationally feasible for global simulations.

Acknowledgments. We thank the LPJmL developers Kirsten Thonicke, Werner von Bloh, and Sibyll Schaphoff at the Potsdam Institute for Climate Impact Research (PIK) for their support in implementing the newly developed method in LPJmL. The first author gratefully acknowledges financial support from the doctoral scholarship program of the Deutsche Bundesstiftung Umwelt (DBU).

The generative AI models GPT-5.6 Sol and Fable 5 were used to shorten the text in Appendix A, suggest approaches to the proof of Theorem 3.9, assist in formulating the Neumann analytical solution presented in Section 4.1, and improve the language and formatting of mathematical expressions throughout the paper. All AI-generated and AI-improved content was reviewed and verified by the authors, who take full responsibility for the correctness and content of the paper.

Appendix A. Derivation of the discrete scheme. In this appendix, the discrete problem (2.9) is derived from the weak formulation of (2.3) using a finite-element method. The derivation proceeds in three steps: the weak formulation is discretized in time, then in space, and the resulting variational problem is converted into the vector-matrix form of Section 2.

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Throughout, the mesh, the data, and the objects $\mathcal{B}$, Q_j^n, F_i^n, F^n, and M_h are as introduced in Section 2.

A.1. The weak formulation. The system (2.3) is understood in the following weak sense, where $k(u)$ abbreviates the function $x \longmapsto k(u(x), x)$.

Weak formulation [21]. Given data $s \in L^\infty(0,T)$ and $e_0 \in L^\infty(0,D)$, find functions $e \in L^2(G)$ and $u \in L^2(0,T;\, H^1(0,D))$ with $u(0,t) = s(t)$ for all $t \in [0,T]$ such that

$$u(x,t) = \beta(e(x,t),x) \quad \text{for almost all } (x,t) \in G$$

and

$$\int_0^T \int_0^D [e\phi_t - k(u)u_x\,\phi_x]\,dx\,dt = -\int_0^D e_0(x)\phi(x,0)\,dx \tag{A.1}$$

for all $\phi \in H^1(G)$ with $\phi(\cdot,T) = 0$ and $\phi(0,\cdot) = 0$ in the trace sense.

A.2. Time discretization. In (A.1), both derivatives are transferred to the test function, owing to the limited regularity of e and u. Since time stepping is a required feature of our method, we do not employ a space-time finite-element method, but instead first restrict the weak formulation to a single time interval, following an idea of Andreucci and Scarpa [3].

Let (e,u) be a solution of the weak formulation, fix a space-only test function $\varphi \in V_0 := \{v \in H^1(0,D) \mid v(0) = 0\}$ and let $0 \le t_1 < t_2 \le T$. For $0 < \epsilon < (t_2 - t_1)/2$ the cutoff function

$$\chi_\epsilon(t) := \min(1,\ (t-t_1)_+/\epsilon,\ (t_2-t)_+/\epsilon)$$

belongs to $H^1(0,T)$ and vanishes at $t = 0$ and $t = T$, so that $\phi(x,t) := \varphi(x)\chi_\epsilon(t)$ is an admissible test function for (A.1). Inserting it and letting $\epsilon \to 0$, the Lebesgue differentiation theorem yields

$$\begin{aligned}\int_0^D e(x,t_1)\varphi(x)\,dx - \int_0^D e(x,t_2)\varphi(x)\,dx \\ = \int_{t_1}^{t_2}\int_0^D k(u(x,t),x)\,u(x,t)_x\,\varphi(x)_x\,dx\,dt\end{aligned} \tag{A.2}$$

for almost all t_1 and t_2.

Different time-stepping schemes are obtained by applying a quadrature rule to the time integral on the right-hand side of (A.2). Choosing $\theta \in [0,1]$, we use the approximation

$$\int_{t_1}^{t_2} f(t)\,dt \approx (t_2 - t_1)[(1-\theta)f(t_1) + \theta f(t_2)],$$

whose cases $\theta = 0$, $\theta = 1$, and $\theta = 1/2$ amount to the *left rectangle*, *right rectangle*, and *trapezoid* rules, respectively. With the time grid of Section 2 and the notation $e^n(x) := e(x,t_n)$ and $u^n(x) := u(x,t_n)$, approximate solutions of the weak formulation can thus be obtained from the following time-stepping problem.

Time discrete problem. Given $e^n \in L^2(0,D)$ and $u^n \in H^1(0,D)$ with $u^n(0) = s^n$ and $u^n(x) = \beta(e^n(x),x)$ almost everywhere, find $e^{n+1} \in L^2(0,D)$ and $u^{n+1} \in H^1(0,D)$ such that

$$\left(\frac{e^{n+1}-e^n}{\Delta t_{n+1}},\ \varphi\right)_{L^2} = -\Big((1-\theta)k(u^n)(u^n)_x + \theta k(u^{n+1})(u^{n+1})_x,\ \varphi_x\Big)_{L^2}$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

for all $\varphi \in V_0$ and, in addition,

$$u^{n+1}(0) = s^{n+1} \quad \text{and} \quad u^{n+1}(x) = \beta(e^{n+1}(x), x)$$

almost everywhere.

A.3. Space discretization. To discretize in space, we adapt a finite-element method for the Stefan problem developed by Elliott [5] to space-dependent physical parameters, the Neumann boundary condition, and general θ-schemes. On the mesh of Section 2, consider the finite-element space

$$V_h := \big\{ v \in C([0, D]) \;\big|\; v|_{T_i} \in P_1 \text{ for } i = 1, \ldots, \kappa \big\},$$

where P_1 denotes the polynomials of degree ≤ 1, together with the test space $V_{0,h} := \{v \in V_h \mid v(0) = 0\}$. A basis of V_h is given by the nodal basis functions $\varphi_0, \ldots, \varphi_\kappa$ determined by $\varphi_i(x_j) = \delta_{ij}$, where δ_{ij} is the Kronecker delta; the functions $\varphi_1, \ldots, \varphi_\kappa$ form a basis of $V_{0,h}$.

The data of Section 2 are connected to the continuous model as follows. The conductivities k_f and k_u are constant on each element with values k_i^f and k_i^u, so that

$$k(u, x) = k_i(u) \quad \text{for all } x \in T_i,$$

with k_i from (2.4). The heat capacities and the latent heat are taken to be piecewise affine, $c_f, c_u, L \in V_h$, with the nodal values c_i^f, c_i^u, and L_i of Section 2 at the nodes $x_1, \ldots, x_\kappa$.

Two modifications are made when restricting the time discrete problem to V_h. First, even for $e \in V_h$ the function $x \longmapsto \beta(e(x), x)$ does not in general belong to V_h; the constitutive relation can therefore not be imposed exactly and is instead enforced at the nodes $x_1, \ldots, x_\kappa$ only. Second, in the time-derivative term, the L^2 inner product is replaced by the quadrature-based product

$$(f, g)_h := \sum_{i=1}^{\kappa} \frac{h_i}{2} [f(x_{i-1})g(x_{i-1}) + f(x_i)g(x_i)], \quad f, g \in V_h, \tag{A.3}$$

which results from applying the trapezoid rule on each element and defines an inner product on V_h. This quadrature produces the diagonal, *lumped*, mass matrix below. We arrive at the following problem.

Discrete problem in variational form. Given $e^n, u^n \in V_h$ with $u^n(x_0) = s^n$ and $u^n(x_i) = \beta(e^n(x_i), x_i)$ for $i = 1, \ldots, \kappa$, find $e^{n+1}, u^{n+1} \in V_h$ such that

$$\left(\frac{e^{n+1} - e^n}{\Delta t_{n+1}},\ \varphi \right)_h = -\Big((1-\theta) k(u^n)(u^n)_x + \theta k(u^{n+1})(u^{n+1})_x,\ \varphi_x \Big)_{L^2} \tag{A.4}$$

holds for all $\varphi \in V_{0,h}$ and, in addition,

$$\text{(i) } u^{n+1}(x_0) = s^{n+1} \quad \text{and} \quad \text{(ii) } u^{n+1}(x_i) = \beta(e^{n+1}(x_i), x_i) \text{ for } i = 1, \ldots, \kappa. \tag{A.5}$$

A.4. Conversion to the vector-matrix form. We express the functions in the nodal basis as

$$u^n = s^n \varphi_0 + \sum_{j=1}^{\kappa} \gamma_j^n \varphi_j \quad \text{and} \quad e^n = \sum_{j=1}^{\kappa} \eta_j^n \varphi_j,$$

ETNA
Kent State University and
Johann Radon Institute (RICAM)

so that $\gamma_j^n = u^n(x_j)$ and $\eta_j^n = e^n(x_j)$ are the nodal values at $x_1, \dots, x_\kappa$ and the boundary condition (i) is built into the representation of u^n. Normalizing the surface value $e^n(x_0)$ to zero entails no loss of generality: it enters neither (A.4) nor the nodal relation (ii). Consequently, no enthalpy unknown is associated with the surface node.

By linearity, (A.4) holds for all $\varphi \in V_{0,h}$ if and only if it holds for the basis functions $\varphi_1, \dots, \varphi_\kappa$. We compute the resulting inner products in turn. For the time-derivative term, (A.3) gives

$$(\varphi_i, \varphi_j)_h = \sum_{k=1}^{\kappa} \frac{h_k}{2} \delta_{ij}(\delta_{i,k-1} + \delta_{i,k}) = \delta_{ij}(M_h)_{ii}, \quad i = 1, \dots, \kappa,\ j = 1, \dots, \kappa,$$

with the mass matrix M_h of Section 2, and hence

$$\left(\frac{e^{n+1} - e^n}{\Delta t_{n+1}},\ \varphi_i \right)_h = \frac{1}{\Delta t_{n+1}} [M_h(\eta^{n+1} - \eta^n)]_i.$$

For the conduction term at time level n, we restrict the integral to a single element. On T_j, we have

$$(\varphi_i)_x = \frac{\delta_{j,i} - \delta_{j,i+1}}{h_j} \quad \text{and} \quad (u^n)_x = \frac{u^n(x_j) - u^n(x_{j-1})}{h_j},$$

where $u^n(x_0) = s^n$ and $u^n(x_i) = \gamma_i^n$ for $i \geq 1$. Since both derivatives are constant on T_j, we obtain

$$\left(k(u^n)(u^n)_x,\ (\varphi_i)_x\right)_{L^2(T_j)} = (\delta_{j,i} - \delta_{j,i+1}) \frac{u^n(x_j) - u^n(x_{j-1})}{h_j} \overline{k(u^n, \cdot)}^{\,j}, \tag{A.6}$$

where $\overline{f}^{\,j}$ denotes the average of a function f over T_j; here $f(x) = k(u^n(x), x)$, which on T_j takes the values k_j^f and k_j^u according to the sign of $u^n(x)$. The central step is to evaluate this average in closed form.

LEMMA A.1. *For $j = 1, \dots, \kappa$,*

$$\frac{u^n(x_j) - u^n(x_{j-1})}{h_j} \overline{k(u^n, \cdot)}^{\,j} = Q_j^n(\gamma^n),$$

with Q_j^n as defined in (2.6).

Proof. Abbreviate the endpoint temperatures $a := u^n(x_{j-1})$ and $b := u^n(x_j)$, so that $a = s^n$ for $j = 1$ and $a = \gamma_{j-1}^n$ for $j \geq 2$, while $b = \gamma_j^n$ in all cases; with these abbreviations, (2.6) reads

$$Q_j^n(\gamma^n) = \frac{k_j(b)b - k_j(a)a}{h_j}.$$

If $a = b$, then u^n is constant on T_j and both sides vanish. So assume $a \neq b$. Then u^n is strictly monotone on T_j and vanishes at most at one point, so the value of the conductivity at $u = 0$ does not affect the average, and we claim that

$$\overline{k(u^n, \cdot)}^{\,j} = k_j(b)\omega + k_j(a)(1 - \omega), \quad \omega := \frac{b}{b - a}. \tag{A.7}$$

Indeed, if a and b have the same strict sign, then u^n has this sign on all of T_j, the average equals $k_j(a) = k_j(b)$, and (A.7) holds trivially. Otherwise, $\operatorname{sgn}(a) \neq \operatorname{sgn}(b)$, and a simple

ETNA
Kent State University and
Johann Radon Institute (RICAM)

geometric consideration shows that the proportion v of the interval $[x_{j-1}, x_j]$ on which $\operatorname{sgn}(u^n) = \operatorname{sgn}(b)$ satisfies

$$\frac{v}{h_j} = \frac{b}{b-a},$$

while $\operatorname{sgn}(u^n) = \operatorname{sgn}(a)$ on the remaining part of measure $h_j - v$. Hence

$$\overline{k(u^n, \cdot)}^{\,j} \cdot h_j = k_j(b) \cdot v + k_j(a) \cdot (h_j - v),$$

and a division by h_j shows (A.7). Substituting (A.7) finally yields

$$\frac{b-a}{h_j}\left[k_j(b)\frac{b}{b-a} + k_j(a)\left(1 - \frac{b}{b-a}\right)\right] = \frac{k_j(b)b - k_j(a)a}{h_j} = Q_j^n(\gamma^n).$$

This completes the proof of the lemma. □

Algorithm 2 Floating-point implementation of the Katzenelson algorithm.

Require: Initial guess x, function $\Phi(\cdot)$, Jacobian $D\Phi(\cdot)$, volumetric latent heat vector L, absolute tolerance t_a, relative tolerance t_r, characteristic enthalpy scale s_x.

$J \leftarrow D\Phi(x)$
$r \leftarrow \Phi(x)$
$r_{\mathrm{ini}} \leftarrow \|r\|$
while $\|r\| > r_{\mathrm{ini}}\mathrm{t}_r + \mathrm{t}_a$ **do**
  $v \leftarrow -J^{-1}\Phi(x)$
  $\Lambda \leftarrow [-x_i/v_i \mid i \in \{1, \ldots, \kappa\},\ |v_i| > \mathrm{t}_r]$
  $\Lambda \leftarrow \operatorname{concat}(\Lambda, [(L_i - x_i)/v_i \mid i \in \{1, \ldots, \kappa\},\ |v_i| > \mathrm{t}_r,\ L_i > s_x\mathrm{t}_r])$
  $\lambda_1 \leftarrow 1$
  $\lambda_2 \leftarrow 1$
  for $\lambda \in \Lambda$ **do**
    if $\lambda > \mathrm{t}_r \wedge \lambda \le \lambda_1$ **then**
      $\lambda_2 \leftarrow \lambda_1$
      $\lambda_1 \leftarrow \lambda$
    else if $\lambda > \mathrm{t}_r \wedge \lambda \le \lambda_2$ **then**
      $\lambda_2 \leftarrow \lambda$
    end if
  end for
  if $|\lambda_1 - \lambda_2| > \mathrm{t}_r \vee \lambda_1 > 1 - \mathrm{t}_r$ **then**
    $J \leftarrow D\Phi\left(x + \dfrac{\lambda_1 + \lambda_2}{2} v\right)$
    $\lambda \leftarrow \lambda_1$
  else
    $x \leftarrow x + (\lambda_1/2)v$
    Draw $\xi = (\xi_1, \ldots, \xi_\kappa)^{\mathrm{T}}$ with $\xi_i \overset{\mathrm{iid}}{\sim} \mathcal{U}(-1/2, 1/2)$
    $x \leftarrow x + s_x\, 10^{-8}\xi$
    $\lambda \leftarrow 0$
  end if
  $x \leftarrow x + \lambda v$
  $r \leftarrow \Phi(x)$
end while

ETNA
Kent State University and
Johann Radon Institute (RICAM)

Summing (A.6) over the elements and applying Lemma A.1 yields

$$\big(k(u^n)(u^n)_x,\ (\varphi_i)_x\big)_{L^2} = \sum_{j=1}^{\kappa} (\delta_{j,i} - \delta_{j,i+1}) Q_j^n(\gamma^n) = F_i^n(\gamma^n),$$

with F_i^n as defined in (2.7). The conduction term at time level $n+1$ is computed analogously and equals $F_i^{n+1}(\gamma^{n+1})$.

Combining the three terms and collecting the equations for $i = 1, \ldots, \kappa$ with the map F^n from (2.8), the variational equation (A.4) is equivalent to

$$\frac{1}{\Delta t_{n+1}} M_h(\eta^{n+1} - \eta^n) = -(1-\theta) F^n(\gamma^n) - \theta F^{n+1}(\gamma^{n+1}).$$

Moreover, since c_f, c_u, and L are piecewise affine, evaluating $\beta(\cdot, x_i)$ at a node involves precisely the nodal data c_i^f, c_i^u, and L_i; the nodal relation (ii) in (A.5) is therefore equivalent to $\gamma^{n+1} = \mathcal{B}(\eta^{n+1})$, and likewise $\gamma^n = \mathcal{B}(\eta^n)$ for the given data at time level n. Substituting both relations turns the system into the discrete problem (2.9) of Section 2. Finally, the initialization $\eta_i^0 = e_0(x_i)$ amounts to interpolating the initial datum at the nodes $x_1, \ldots, x_\kappa$.

Appendix B. Pseudocode for the implementation of the Katzenelson algorithm. The pseudocode is listed on the previous page. The notation $[\,\cdot\,]$ denotes a finite list. The concatenation of lists is denoted by the concat operator.

For soil heat transfer, the parameters $\mathrm{t}_r = 10^{-12}$, $\mathrm{t}_a = 10^{-6}$, and $s_x = 10^6$ performed well, and were used for most of the simulations.

REFERENCES

[1] J. Aalto, O. Karjalainen, J. Hjort, and M. Luoto, *Statistical forecasting of current and future circum-Arctic ground temperatures and active layer thickness*, Geophys. Res. Lett., 45 (2018), pp. 4889–4898.
[2] V. Alexiades and A. Solomon, *Mathematical Modeling of Melting and Freezing Processes*, Hemisphere, Washington, 1993.
[3] D. Andreucci, *Lecture Notes on the Stefan Problem*, Lecture Notes, 2005 (Accessed: 24 March 2024). http://www.sbai.uniroma1.it/pubblicazioni/doc/phd_quaderni/02-1-and.pdf
[4] J. Crank, *Free and Moving Boundary Problems*, Clarendon Press, Oxford University Press, New York, 1987.
[5] C. M. Elliott, *Error analysis of the enthalpy method for the Stefan problem*, IMA J. Numer. Anal., 7 (1987), pp. 61–71.
[6] A. Friedman, *The Stefan problem in several space variables*, Trans. Amer. Math. Soc., 133 (1968), pp. 51–87.
[7] T. Fujisawa and E. S. Kuh, *Piecewise-linear theory of nonlinear networks*, SIAM J. Appl. Math., 22 (1972), pp. 307–328.
[8] F. P. Gantmacher and M. G. Krein, *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, revised ed., AMS Chelsea Publishing, Providence, 2002.
[9] J. Katzenelson, *An algorithm for solving nonlinear resistor networks*, Bell System Tech. J., 44 (1965), pp. 1605–1620.
[10] S. Lange, M. Mengel, S. Treu, and M. Büchner, *ISIMIP3a atmospheric climate input data (v1.0)*, ISIMIP Repository, 2022. DOI: https://doi.org/10.48364/ISIMIP.982724
[11] D. Lawrence et al., *Technical description of version 5.0 of the Community Land Model (CLM)*, Tech. Report, National Center for Atmospheric Research (NCAR), Boulder, 2018.
[12] G. H. Meyer, *Multidimensional Stefan problems*, SIAM J. Numer. Anal., 10 (1973), pp. 522–538.
[13] D. J. Nicolsky, V. E. Romanovsky, V. A. Alexeev, and D. M. Lawrence, *Improved modeling of permafrost dynamics in a GCM land-surface scheme*, Geophys. Res. Lett., 34 (2007). DOI: https://doi.org/10.1029/2007GL029525
[14] G.-Y. Niu, Z.-L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, A. Kumar, K. Manning, D. Niyogi, E. Rosero, M. Tewari, and Y. Xia, *The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements*, J. Geophys. Res., 116 (2011). DOI: https://doi.org/10.1029/2010JD015139

ETNA
Kent State University and
Johann Radon Institute (RICAM)

[15] J. OBU, *How much of the Earth's surface is underlain by permafrost?*, J. Geophys. Res.: Earth Surface, 126 (2021), Paper No. e2021JF006123, 5 pages.
DOI: https://doi.org/10.1029/2021JF006123

[16] S. SCHOLTES, *Introduction to Piecewise Differentiable Equations*, Springer, New York, 2012.

[17] D. A. TARZIA, *Relationship between Neumann solutions for two-phase Lame–Clapeyron–Stefan problems with convective and temperature boundary conditions*, Thermal Sci., 21 (2017), pp. 187–197.

[18] M. J. TODD, *An introduction to piecewise-linear homotopy algorithms for solving systems of equations*, in Topics in Numerical Analysis (Lancaster, 1981), P. R. Turner, ed., Lecture Notes in Math., 965, Springer, Berlin, 1982, pp. 149–202.

[19] ———, *On the computational complexity of piecewise-linear homotopy algorithms*, Math. Programming, 24 (1982), pp. 216–224.

[20] N. TUBINI, S. GRUBER, AND R. RIGON, *A method for solving heat transfer with phase change in ice or soil that allows for large time steps while guaranteeing energy conservation*, The Cryosphere, 15 (2021), pp. 2541–2568.

[21] A. VISINTIN, *Introduction to Stefan-type problems*, in Handbook of Differential Equations: Evolutionary Equations. Vol. IV, C. M. Dafermos and M. Pokorný, eds., Handb. Differ. Equ., Elsevier, Amsterdam, 2008, pp. 377–484.

[22] V. VOLLER AND M. CROSS, *Accurate solutions of moving boundary problems using the enthalpy method*, Int. J. Heat Mass Transfer, 24 (1981), pp. 545–556.

[23] V. R. VOLLER, *Implicit finite-difference solutions of the enthalpy formulation of Stefan problems*, IMA J. Numer. Anal., 5 (1985), pp. 201–214.

[24] W. VON BLOH, S. SCHAPHOFF, C. MÜLLER, S. ROLINSKI, K. WAHA, AND S. ZAEHLE, *Implementing the nitrogen cycle into the dynamic global vegetation, hydrology, and crop growth model LPJmL (version 5.0)*, Geosci. Model. Develop., 11 (2018), pp. 2789–2812.

[25] R. WANIA, I. ROSS, AND I. C. PRENTICE, *Integrating peatlands and permafrost into a dynamic global vegetation model: 1. Evaluation and sensitivity of physical land surface processes*, Global Biogeochem. Cycles, 23 (2009). DOI: https://doi.org/10.1029/2008GB003412

[26] R. E. WHITE, *An enthalpy formulation of the Stefan problem*, SIAM J. Numer. Anal., 19 (1982), pp. 1129–1157.

[27] ———, *A numerical solution of the enthalpy formulation of the Stefan problem*, SIAM J. Numer. Anal., 19 (1982), pp. 1158–1172.

[28] Y. ZHANG, S. K. CAREY, AND W. L. QUINTON, *Evaluation of the algorithms and parameterizations for ground thawing and freezing simulation in permafrost regions*, J. Geophys. Res.: Atmospheres, 113 (2008). DOI: https://doi.org/10.1029/2007JD009343

[29] W. ZULEHNER, *Numerische Mathematik: eine Einführung anhand von Differentialgleichungsproblemen. Band 1. Stationäre Probleme*, Birkhäuser, Basel, 2008.

[30] ———, *Numerische Mathematik: eine Einführung anhand von Differentialgleichungsproblemen. Band 2. Instationäre Probleme*, Birkhäuser, Basel, 2011.