Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The ballon model revisited

John Gore

john.gore@vanderbilt.edu

Vanderbilt University Medical Center https://orcid.org/0000-0002-9480-0900

Tzu-wei Lee

Vanderbilt University

Yang Jiaao

Arizona State University

Abbe L. Knutsson

Vanderbilt University

warren boschen

Vanderbilt University

Muwei Li

Vanderbilt University Medical Center

Kurt G. Schilling

Vanderbilt University Medical Center

Adam W. Anderson

Vanderbilt University

Zhaohua Ding

Vanderbilt University

Yurui Gao

Vanderbilt University

Research Article

Keywords: Hemodynamic response function, BOLD, fMRI, balloon model

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11197888/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

The balloon model revisited

John C. Gore[1,2,3*], Tzu-Wei Li[1,2], Yang Jiao[4], Abbe L. Knutsson[1], Warren W. Boschen[1,2], Muwei Li[1,3], Kurt Schilling[1,3], Adam W. Anderson[1,2,3], Zhaohua Ding[1,5] and Yurui Gao[1,2]

[1]Vanderbilt University Institute of Imaging Science, Vanderbilt University, Nashville, TN 37232, USA

[2]Department of Biomedical Engineering, Vanderbilt University, Nashville, TN 37232, USA

[3]Department of Radiology and Radiological Sciences, Vanderbilt University Medical Center, Nashville, TN 37232, USA

[4]Materials Science and Engineering, Arizona State University; Tempe, 85287, USA

[5]Department of Electrical and Computer Engineering, Vanderbilt University, Nashville, TN 37232, USA

*Corresponding author. Email: john.gore@vanderbilt.edu

A closed-form hemodynamic response function links the initial dip, peak and undershoot of BOLD fMRI to vascular physiology.

Abstract

Functional MRI based on the blood oxygenation level–dependent (BOLD) effect infers neural activity from hemodynamic changes through a hemodynamic response function (HRF). We revisit the balloon model of the BOLD effect and show that Grubb's empirical flow–volume power law can be used to determine how vascular transit time varies with flow and is governed by a single index of venous distensibility. Linearizing the resulting equations yields an explicit closed-form HRF, expressed in physiologically recognizable parameters, that reproduces the canonical response. The initial dip, time to peak and poststimulus undershoot are each traced to identifiable terms, and the relations assumed in calibrated BOLD signals emerge as outputs. When oxygen consumption is specified independently of flow, as in earlier generalizations, the analytic solution quantifies the specific relationships between the latency and depth of an initial dip and a neurovascular delay. The new model will prove useful for explaining variations in HRF in white matter, across the brain, and between different populations.

Introduction

We reconsider the model and equations of Buxton et al. (1), who described the blood oxygenation-level dependent (BOLD) signal (2), which is the basis for functional MRI of the brain and spine. FMRI has become well established as the premier noninvasive imaging technology for mapping the functional organization of the brain. FMRI acquisitions are used to detect localized activations caused by changes in a stimulus condition, to detect transient signals evoked by very short stimuli (event-related fMRI) and to detect fluctuations in the baseline signals in trains of images obtained while subjects are in a "resting state" (3). Correlations between such fluctuations are deemed to depict functional connectivity within and between networks.

Although data-driven methods for analyzing sets of fMRI images, such as independent component analysis (ICA), have been used in various applications, the most common approach for deriving measures of brain activity involves the use of a general linear model that assumes a specific relationship between a transient change in neural activity and the BOLD response, called the hemodynamic response function (HRF) (4, 5). There is a growing body of literature on the appropriate design and selection of HRF for optimizing sensitivity to brain activity (6, 7), although rather less emphasis has been placed on mechanistic models that predict the form of HRF. Analytical treatments for hemodynamic models have been proposed previously. Friston et al. (4) placed the balloon model in state-space form and derived Volterra kernels, the first-order kernel being an impulse-response representation of the HRF; Robinson, Aquino, Pang and colleagues linearized spatially extended, poroelastic models of cortical hemodynamics, formed their transfer functions, identified the poles and inverted them to obtain closed-form responses and response modes (8–12). The present work follows a similar route (linearize, transform and

57 invert) for a lumped model, with the advantage that the resulting expressions are compact and
58 that each term corresponds to an identifiable physiological quantity.

59 BOLD signals from neural tissues are indirectly related to underlying neural metabolic activity
60 (13, 14), and the nature of that coupling profoundly affects the temporal response of the BOLD
61 signal to transient changes in conditions. Understanding the factors that govern HRF is key to
62 interpreting the variations in HRF that are evident across the brain, between individuals, and, in
63 particular, between gray and white matter (15–17). Moreover, previous investigators have
64 attempted to calibrate the BOLD effect in terms of underlying metabolic and physiological
65 properties (18–20) and have relied on concepts derived from the balloon model and empirical
66 observations for validation. Here, we revisit the assumptions of previous models and rewrite
67 their constitutive equations to derive a more complete theoretical explanation of the BOLD effect
68 and HRF and quantify how specific features are related to underlying, measurable physiological
69 parameters. We first describe the original model of Buxton et al. (1) as modified by Obata (21)
70 and identify the conditions under which the BOLD signal is linearly proportional to the
71 deoxyhemoglobin content. We then recast Grubb's empirical flow–volume power law (22) in
72 terms of vascular transit times and the mechanical properties of the draining vessels. Using these
73 new insights, we explain how the exact shape of the HRF is governed by specific factors.
74 Following earlier versions of the model that treat oxygen consumption as an input independent of
75 flow (21, 23), we then remove the assumption that oxygen consumption is determined by flow,
76 which permits increases in metabolism to occur before changes in flow and cause an initial dip,
77 and derive closed-form expressions for that dip. We finally solve the linearized system exactly,
78 first from oxygen consumption to signal and then for the complete response to a neural event.
79 Numerical results are then presented, including predictions of the HRF for different underlying

parameters and fits of the theoretical functions to the canonical HRF used in standard analysis software. These fits are intended as a practical demonstration that the model reproduces the standard response with physiologically plausible parameters.

The symbols used are defined at first use and are listed in Table S1.

Results

(a) The balloon model of Buxton et al. was modified from Obata

The original balloon model (1) treats the postcapillary venous compartment in tissue as an expandable balloon, fed by an inflow $F_{\text{in}}(t)$ and drained by an outflow $F_{\text{out}}(t)$. Two time-varying quantities of relevance are the venous blood volume $V(t)$ and the total quantity of deoxyhemoglobin it contains $Q(t)$. Conservation of blood volume and deoxyhemoglobin levels

$$\frac{dV}{dt} = F_{\text{in}}(t) - F_{\text{out}}(t) \tag{1}$$

$$\frac{dQ}{dt} = F_{\text{in}}(t)\, C_a\, E(t) - F_{\text{out}}(t)\, \frac{Q}{V} \tag{2}$$

where C_a is the arterial oxygen content, $E(t)$ is the net oxygen extraction fraction, and Q/V is the venous deoxyhemoglobin concentration. It is convenient to normalize each quantity by its resting value, written as $v = V/V_0$, $q = Q/Q_0$, $f_{\text{in}} = F_{\text{in}}/F_0$ and $f_{\text{out}} = F_{\text{out}}/F_0$, all of which are unity at rest.

The mean transit time of blood through the compartment isτ_0

$$\tau_0 = \frac{V_0}{F_0} \tag{3}$$

Equations (1) and (2) become

$$\tau_0\, \frac{dv}{dt} = f_{\text{in}}(t) - f_{\text{out}}(t) \tag{4}$$

$$\tau_0 \frac{dq}{dt} = f_{\text{in}} \frac{E(f_{\text{in}})}{E_0} - f_{\text{out}} \frac{q}{v} \quad (5)$$

where E_0 is the resting oxygen extraction fraction. Two other relationships are relevant. The first is the relation between oxygen extraction and flow. Buxton et al. assumed that oxygen delivery is flow-limited with a fixed capillary permeability–surface area product, which gives

$$E(f) = 1 - (1 - E_0)^{1/f} \quad (6)$$

Therefore, the extraction decreases as the flow increases. The second limitation is the relationship between venous outflow and venous volume. Buxton et al. adopted the empirical steady-state power law measured by Grubb et al. (22),

$$f_{\text{out}}(v) = v^{1/\alpha} \quad (7)$$

with $\alpha \approx 0.38$, so that $1/\alpha \approx 2.63$. Equations (4)–(7) provide a complete description of how neural activity drives f_{in}. Several variants of that relationship have been used; below, we adopt the widely used second-order form of Friston et al. (4), which is given in eq. 25 below (24, 25).

(b) The signaling equation relating ΔR_2^* to deoxyhemoglobin

The BOLD signal changes mainly because of changes in the transverse relaxation rate of tissue magnetization. For a gradient-echo acquisition at echo time TE, the signal from a voxel is

$$S(\text{TE}) = S_0 \, e^{-R_2^* \, \text{TE}} \quad (8)$$

such that a change ΔR_2^* produced by activation results in a fractional signal change

$$\frac{\Delta S}{S} = e^{-\Delta R_2^* \text{TE}} - 1 \simeq -\Delta R_2^* \, \text{TE} \qquad (|\Delta R_2^* \, \text{TE}| \ll 1) \quad (9)$$

The linear approximation in (9) is well satisfied in practice. At 3T with TE = 30 ms, a 1% signal change corresponds to $\Delta R_2^* \approx -0.33$ s^{-1} $-\Delta R_2^* \text{TE} \approx 0.01$. ΔR_2^* depends on separate extravascular and intravascular contributions. In the static dephasing regime around a random distribution of vessels larger than the diffusion length, the extravascular relaxation rate is

proportional to the volume fraction of deoxygenated vessels multiplied by the resonant frequency offset they produce (26–28), while the total deoxyhemoglobin content is proportional to the product of venous volume and extraction:

$$\Delta R_{2,E}^{*} = 4.3\,\Delta\omega_0 E_0 V_0\,(q-1) \tag{10}$$

Here, $\Delta\omega_0 = \gamma_H\,\frac{4}{3}\pi\,\Delta\chi_0\,\mathrm{Hct}\,B_0$ is the frequency offset at the surface of a fully deoxygenated vessel with hematocrit Hct, γ_H is the proton gyromagnetic ratio, and the numerical coefficient follows from Monte Carlo and analytic treatments of randomly oriented cylinders (27, 28).

The intravascular contribution arises because the transverse relaxation rate of blood itself depends on its oxygenation and varies approximately linearly with extraction over the physiological range. The venous deoxyhemoglobin concentration is q/v

$$\Delta R_{2,I}^{*} = r_0 E_0 V_0 \left(\frac{q}{v} - 1\right) \tag{11}$$

where r_0 is the slope of the intravascular relaxation rate against extraction. A third term arises because blood and tissue do not have the same signal per unit volume: when the venous volume increases, blood displaces tissue, and if ε_r is the ratio of intravascular to extravascular signal at rest, the net signal changes in proportion to $(1-\varepsilon_r)(1-v)$.

Collecting these contributions and applying (9) gives the signal equation in the form derived by Obata et al. (21):

$$\frac{\Delta S}{S} = V_0\left[k_1(1-q) + k_2\left(1-\frac{q}{v}\right) + k_3(1-v)\right] \tag{12}$$

$$k_1 = 4.3\,\Delta\omega_0 E_0\,\mathrm{TE}, \qquad k_2 = \varepsilon_r r_0 E_0\,\mathrm{TE}, \qquad k_3 = 1-\varepsilon_r \tag{13}$$

Obata's revision of the original coefficients separately revealed three physical contributions: k_1 multiplied the *total* deoxyhemoglobin content and was extravascular; k_2 multiplied the deoxyhemoglobin *concentration* and was intravascular; and k_3 multiplied the *volume* and

represented the change in S_0 caused by blood displacing tissue. At 3 T with TE = 30 ms, $E_0 = 0.34$, $\varepsilon_r = 1.43$, $r_0 \approx 25$ s^{-1} and $\Delta\omega_0 \approx 80.6$ s^{-1}, which give $k_1 = 3.535$, $k_2 = 0.365$ and $k_3 = -0.430$. The constants depend on the field strength and echo time (29). Later sections show that quantities used in calibrated BOLD signals follow from these three constants alone.

(c) Blood flow, blood volume and transit time

The key to interpreting BOLD signal changes is the relationship between blood volume and blood flow. Writing CBF for cerebral blood flow and CBV for cerebral blood volume, with subscript 0 denoting the resting value throughout, the relation measured by Grubb et al. (22) during hypercapnia is

$$\frac{\mathrm{CBV}}{\mathrm{CBV}_0} = \left(\frac{\mathrm{CBF}}{\mathrm{CBF}_0}\right)^{\alpha}, \qquad \alpha \approx 0.38 \tag{14}$$

where α is Grubb's exponent, a dimensionless constant.

Separately, and seemingly unrelated, the Stewart–Hamilton central volume principle states that for any compartment at steady state, the volume V, the flow F through it, and the mean transit time τ of a tracer are related:

$$V = F\,\tau \tag{15}$$

Here, τ, the mean transit time (MTT), is the average time a small element of blood spends in the compartment in seconds. Equation (15) says that volume and flow are related by a single quantity with units of time, while (14) says that they are related by a dimensionless exponent. If τ is a fixed property of the vasculature, (15) will force $\alpha = 1$, and Grubb's measured $\alpha = 0.38$ will contradict this.

τ is not a constant for resolving this apparent paradox. It is a variable property of the venous compartment and is determined by the instantaneous pressure–volume relationship. Dividing (15) for an activated state by the conditions at rest,

$$\frac{\tau}{\tau_0} = \frac{V/V_0}{F/F_0} = \frac{v}{f} \tag{16}$$

and substituting Grubb's relation (14),

$$\tau(f) = \tau_0 f^{\alpha-1} \qquad \text{or} \qquad \tau(v) = \tau_0 v^{1-1/\alpha} \tag{17}$$

Grubb's law and the central volume principle are simultaneously true but describe different properties. Equation (17) shows that Grubb's exponent reflects the rate at which transit time decreases under vasodilation and increased flow. With $\alpha = 0.38$, $\tau \propto f^{-0.62}$: a 60% flow increase shortens the MTT by 25%.

Considering two limits is instructive. First, $\alpha = 1$ would mean that transit time does not change with changing flow, corresponding to a rigid tube of fixed length whose volume tracks flow exactly, which is unphysical for a venous system of limited compliance. Second, $\alpha = 0$ would mean that the volume is fixed and that all of the flow change is accommodated by an increase in velocity. The measured values are between these values.

Combining equations (4) and (16), the outflow can be written as $f_{\text{out}} = v\,\tau_0/\tau(v)$. The model requires specifying the transit-time function $\tau(v)$, and Grubb's law is one choice of $\tau(v)$.

(d) Mechanical properties of the vasculature

For laminar flow through a cylindrical venule of radius r and fixed length L, Poiseuille's law gives $F \propto \Delta P\, r^4$, where ΔP is the pressure drop that drives flow along the vessel, while the volume satisfies $V \propto r^2 L$. Hence, the flow relative to rest is

$$f_{\text{out}} \propto \frac{\Delta P}{\Delta P_0}\, v^2 \tag{18}$$

If the draining vascular bed is distensible and the downstream pressure is fixed, the driving pressure increases as the compartment inflates, which we write as $\Delta P/\Delta P_0 = v^{\beta}$, where $\beta \geq 0$ is a dimensionless stiffness index ($\beta = 0$ describes a pressure-clamped vessel and a larger β a stiffer vessel). Then,

$$f_{\text{out}}(v) = v^{\,2+\beta} \qquad \text{and} \qquad \tau(v) = \tau_0\, v^{-(1+\beta)} \tag{19}$$

Comparing (19) with (17) identifies Grubb's constant as a derived, rather than fundamental, quantity:

$$\alpha = \frac{1}{2+\beta} \tag{20}$$

For example, $\beta = 0$ gives $\alpha = 0.50$; Grubb's $\alpha = 0.38$ corresponds to $\beta = 0.63$; and a venous-specific $\alpha \approx 0.23$ (30) corresponds to $\beta \approx 2.3$. The measured range of α corresponds to a physically reasonable range of venous stiffness.

(e) An equivalent linear windkessel

An alternative viewpoint considers the vasculature to have linear compliance C (volume change per unit pressure) and a fixed downstream resistance R that does not change with activation. If p is the intracompartment venous pressure, p_0 is the resting value and p_v is the outflow pressure, then $F_{\text{out}} = (p - p_v)/R$, so $f_{\text{out}} = (p - p_v)/(p_0 - p_v)$. The compliance relation is $v - 1 = C(p - p_0)/V_0$. Eliminating p,

$$f_{\text{out}}(v) = 1 + g\,(v - 1), \qquad g = \frac{V_0}{C\,(p_0 - p_v)} \tag{21}$$

where g is the normalized venous stiffness. Linearizing (19) about $v = 1$ gives $g = 2 + \beta = 1/\alpha$, so Grubb's constant is considered to be the reciprocal of the normalized stiffness of the draining bed.

(f) Delayed compliance

The HRF is observed to possess a significant poststimulus undershoot that outlasts the flow response (31, 32), a feature that is not possible for a purely elastic balloon. Following Mandeville et al. (33), a viscous term may be added so that the outflow lag volume, with a viscoelastic time constant τ_v in seconds, is

$$f_{\text{out}} = v^{2+\beta} + \tau_v \frac{dv}{dt} \tag{22}$$

which, combined with the balance equation (4), gives

$$\frac{dv}{dt} = \frac{f_{\text{in}} - v^{2+\beta}}{\tau_0 + \tau_v}, \qquad f_{\text{out}} = f_{\text{in}} - \tau_0 \frac{dv}{dt} \tag{23}$$

τ_v varies between inflation (τ_{in}, used when $f_{\text{in}} > v^{2+\beta}$) and deflation ($\tau_{\text{de}}$). Expanding to the first order about the resting point, writing $v = 1 + \delta v$ and $f_{\text{in}} = 1 + \delta f$

$$v^{2+\beta} = (1 + \delta v)^{2+\beta} \approx 1 + (2 + \beta)\,\delta v$$

Substituting into (23), $\frac{d\,\delta v}{dt} = \frac{\delta f - (2+\beta)\,\delta v}{\tau_0 + \tau_v} = -\frac{\delta v}{\tau_c} + \frac{\delta f}{\tau_0 + \tau_v}$

which is a first-order rate equation with a volume change time constant

$$\tau_c = \frac{\tau_0 + \tau_v}{2+\beta} = \alpha\,(\tau_0 + \tau_v) \tag{24}$$

using $\alpha = 1/(2 + \beta)$ from (20).

This is another indication of what α controls the dynamic response of the vasculature.

Relation to previous work

The idea that the flow–volume relation should be derived from vessel mechanics rather than imposed as an empirical power law is not new, and the above discussion is a simplification and consolidation of previous literature. Mandeville et al. (33) introduced a postarteriole windkessel with delayed compliance to account for the temporal mismatch they measured between CBV and CBF in the rat cortex, and equation (22) was derived from these data. Kong et al. (34) and Zheng and Mayhew (35) developed viscoelastic windkessel models in which the steady-state Grubb power law and a nonlinear pressure–volume curve are imposed together, with the pressure–volume parameter allowed to vary dynamically so that the vessels exhibit hysteresis between dilation and contraction; Zheng and Mayhew parameterized their model by the inverse Grubb exponent, which is the quantity denoted $g$in (21). Boas et al. (36) computed flow and volume directly from Poiseuille's law over an anatomically realistic vascular network. Resistance scales as r^4, while volume scales as r^2, which underlies (18), is standard in considerations of cerebrovascular physiology.

Here, we combine these suggestions into a simple formulation that shows how the single distensibility exponent β determines Grubb's exponent through (20); that Grubb's law is equivalent to a transit-time relation (17); and that we use these substitutions in support of the theoretical results below. We are not aware of these identities having been written down explicitly before, but the underlying ideas were previously proposed (33–36). Closed-form transfer-function solutions of linearized hemodynamic models have also been previously derived (4, 8–12), but this contribution is the first to obtain them for a lumped model whose parameters are directly physiological.

(g) Revised balloon model equations

Collecting the results above and introducing $u(t)$ as a transient neural excitation (a unit impulse at $t = 0$) stimulating the BOLD response, the revised model becomes

$$\frac{dz}{dt} = \varepsilon\, u(t) - \kappa\, z - \gamma\,(f_{\text{in}} - 1), \qquad \frac{df_{\text{in}}}{dt} = z \tag{25}$$

$$\frac{dv}{dt} = \frac{f_{\text{in}} - v^{\,2+\beta}}{\tau_0 + \tau_v} \tag{26}$$

$$f_{\text{out}} = f_{\text{in}} - \tau_0\, \frac{dv}{dt} \tag{27}$$

$$\tau_0\, \frac{dq}{dt} = f_{\text{in}}\, \frac{E(f_{\text{in}})}{E_0} - f_{\text{out}}\, \frac{q}{v} \tag{28}$$

together with the signal equation (12) and constants (13). In (25), which is the second-order flow-driving stage of Friston et al. (4), z is a vasoactive signal generated by neural activity with efficacy ε, decaying at rate κ (equivalently with time constant $\tau_z = 1/\kappa$) and subject to autoregulatory feedback of strength γ (in s^{-2}; for comparison with earlier work, we also report $\tau_f \equiv 1/\gamma$, which has units of s^2); it is an intermediate quantity that converts a brief neural event into a smooth flow response, but only f_{in} is observable. The symbol z is used here in place of s in the original formulation so that s may be reserved for the Laplace variable below. The residual system parameters are τ_0, the resting mean transit time in seconds; β, a dimensionless stiffness index that relates volume to pressure; and τ_{in} and τ_{de}, the viscoelastic time constants.

Equations (25)–(28) can be integrated numerically to predict the HRF for any set of input parameters, and this process is demonstrated in the Results section. First, we extend the model to explicitly introduce the contributions of changes in oxygen metabolism and solve the linearized system in closed form.

(h) Oxygen metabolism as an independent drive

Optical and MRI measurements have suggested that a positive BOLD response may be preceded by brief negative deflection, the "initial dip", attributed to oxygen consumption preceding the flow response (37, 38); moreover, the existence and magnitude of this response remain contested in gray matter (39, 40), but this response appears prominently in white matter responses (16, 41). The balloon model in its original form can produce an early negative deflection (1), but because extraction is tied to flow, it cannot produce a dip caused by an increase in metabolism that precedes the flow response. The deoxyhemoglobin change in (28) is given by $f\,E(f)/E_0$. The cerebral metabolic rate of oxygen ($CMRO_2$) and the oxygen extraction fraction (OEF) are related as

$$\mathrm{CMRO_2} = \mathrm{CBF} \cdot \mathrm{OEF} \cdot C_a \tag{29}$$

so that, in normalized form, $m = \mathrm{CMRO_2}/\mathrm{CMRO_{2,0}}$ and $e = \mathrm{OEF}/\mathrm{OEF_0}$, $m = fe$. The extraction term in (28) is therefore, by definition, normalized oxygen consumption:

$$m(t) = f\,\frac{E(f)}{E_0} \tag{30}$$

Oxygen consumption is thereby dependent on flow. Extraction is a function of flow alone, so changes in metabolism cannot precede changes in flow, and the coupling between them is not a free parameter but is fixed by the resting extraction fraction. Any initial dip the original model produces is therefore a vascular phenomenon only, arising because the balloon must inflate before its outflow can rise so that washout of deoxyhemoglobin lags an increase in its amount. This mismatch is not a metabolic mismatch that explains this dip. The conditions under which this vascular dip occurs are given below (eqs. 55 and 56).

Allowing oxygen consumption to vary independently of flow has been considered in other analyses. Obata et al. (21) generalized the balloon model by allowing extraction to vary independently in time so that different relative time courses of CBF, CBV and $CMRO_2$ produce different BOLD transients even when the flow responses are similar. Buxton et al. (23) treated $CMRO_2$ as an independent input; Simon and Buxton (42) examined the dynamic dissociation between flow and BOLD; and Tang et al. (43) showed with balloon-model analyses of V1 that transient features, including differences between activations and deactivations, arise within this broader family of models. Here, we adopt the same decoupling method, producing the new results below, which include closed-form expressions for the dip (eqs. 40 and 41); an initial slope that is independent of coupling, stiffness and viscoelasticity (eq. 54); and criteria for when a dip occurs (eqs. 56 and 57).

Allowing $m(t)$ to be independent of flow allows for more varied and realistic HRFs. There is a balance between the production of deoxyhemoglobin and the consumption of oxygen, and (28) becomes:

$$\tau_0 \frac{dq}{dt} = m(t) - f_{\text{out}} \frac{q}{v} \tag{31}$$

Oxygen extraction is then determined by conditions such that

$$e(t) \equiv \frac{\text{OEF}}{\text{OEF}_0} = \frac{m(t)}{f_{\text{in}}(t)} \tag{32}$$

Equation (31) differs from (28) only in transient conditions, but there are two consequences to emphasize before further considering the overall HRF.

(i) Calibrated fMRI and the Davis relation

The venous deoxyhemoglobin concentration is proportional to OEF, so the total deoxyhemoglobin content in the steady state is $q = ve$, and $\delta q = \delta v + \delta e$. Linearizing the

signal equation (12) about the baseline gives $\delta S/S = -(k_1 + k_2)\delta q + (k_2 - k_3)\delta v$; substituting $\delta q = \delta v + \delta e$ and $\delta v = \alpha\, \delta f$,

$$\frac{\delta S}{S} = -M[\alpha\, \delta f + \beta_D\, \delta e], \qquad M = V_0(k_1 + k_3), \qquad \beta_D = \frac{k_1+k_2}{k_1+k_3} \tag{33}$$

M is the calibration constant of Davis et al. (18) — the maximum attainable BOLD change — and β_D their field-dependent exponent. The BOLD signal is driven by a decrease in oxygen extraction, with a smaller opposing term α δf from volume dilation. Using $\delta e = \delta m - \delta f$,

$$\frac{\delta S}{S} = M[(\beta_D - \alpha)\, \delta f - \beta_D\, \delta m] \tag{34}$$

Calibrated BOLD studies report measurements of the steady-state CBF–$CMRO_2$ coupling ratio. This parameter describes the relation between a fractional change in flow and a fractional change in oxygen consumption $n \equiv \delta f/\delta m$; calibrated BOLD studies report values of 2–3 (19, 20). Using this simplifies to (34):

$$\frac{\delta S}{S} = M\, G\, \delta m, \qquad G = n(\beta_D - \alpha) - \beta_D \tag{35}$$

G represents an overall "metabolic gain", which vanishes at $n_{\text{crit}} = \beta_D/(\beta_D - \alpha)$; below that coupling ratio, an increase in $CMRO_2$ produces a *negative* BOLD change. The sign of the metabolic contribution to resting-state BOLD signals therefore depends on whether n (the fractional change in flow per fractional change in oxygen consumption) exceeds a threshold that is itself set by α and hence by venous stiffness.

(j) Dual pathways for stimulating a response

Neural activity can now be considered to drive two pathways that cause changes in BOLD signals: a prompt metabolic response that alters oxygen consumption and a flow response that may lag behind it by a delay Δ.

The metabolic pathway depends on $m(t)$, the time course of the normalized rate of oxygen consumption. Neural activity does not increase oxygen consumption instantaneously. The firing of neurons uses energy supplied by increased fluxes of metabolic intermediates that involve several steps, each with its own kinetics, and for our purposes, only the aggregate effect occurs (38). We therefore describe the metabolic response to an instantaneous impulse via a simple filter constructed to match the minimal properties that are needed — the oxygen consumption starts at rest, rises in a finite time to a peak, and returns to baseline. The simplest first-order stretching of an impulse is a linear filter in which, if the quantity of interest is instantaneously displaced from its equilibrium value and returns toward it at a rate proportional to the displacement, an impulse input produces a single exponentially decaying response. The impulse response of one such stage is therefore

$$k(t) = \frac{1}{\tau_m} e^{-t/\tau_m}, \qquad t \geq 0$$

τ_m = the time constant of recovery, and the response is normalized so that $\int_0^\infty k\,(t)\,dt = 1$. A single stage delays and smooths its input but does not change its area. However, a single relaxation alone increases discontinuously at $t = 0$ and decays from there, so oxygen consumption would have to change infinitely quickly at stimulus onset, which is unphysical. However, two such filters are arranged in series so that the output of the first drives the second, producing a response that matches our requirements. Writing τ_m for the time constant of each filter and a_m for the gain and taking $u(t)$ to be a unit impulse — an idealized instantaneous burst of neural activity — the result is given by the convolution of the two filter impulse responses,

$$k_2(t) = \int_0^t \frac{1}{\tau_m} e^{-(t-t\prime)/\tau_m} \cdot \frac{1}{\tau_m} e^{-t\prime/\tau_m}\, dt'$$

so that

$$\delta m(t) = \frac{a_m}{\tau_m^2}\, t\, e^{-t/\tau_m} \tag{36}$$

The response (36) is zero at t = 0, rises to a peak at $t = \tau_m$ and has an area $\int_0^\infty \delta\, m\, dt = a_m$; thus, a_m is the time-integrated fractional increase in oxygen consumption in units of seconds. A value of T s means that the extra oxygen consumed over the whole response equals T seconds' worth consuming at the resting rate. More generally, for k successive stages, each time constant τ, is given $m(t) = t^{k-1}e^{-t/\tau}/[\tau^k(k-1)!]$, the gamma density of shape k and the scale τ—the same family of functions conventionally used to describe the measured HRF itself (5). Equation (36) is the case in which $k = 2$; the form used for numerical integration is given in *Materials and Methods*.

The flow-inducing signal of (25) can be generated in the same way: in both pathways, a brief neural event is converted into a smooth, delayed physiological response by a second-order linear filter, and the two differ only in their time constants and onset delay Δ. The flow pathway is shown in (25) with the drive delay:

$$\frac{dz}{dt} = \varepsilon\, u(t-\Delta) - \kappa\, z - \gamma\,(f_{\text{in}} - 1), \qquad \frac{df_{\text{in}}}{dt} = z \tag{37}$$

(k) Analytic description of the negative dip

Before considering the overall HRF, a useful measure of the negative dip follows from the equations above. Between $0 \le t \le \Delta$, the flow is not increased, so $f_{\text{in}} = 1$; the venous volume is constant, $v = 1$, and $dv/dt = 0$. In this interval tThe balloon was not allowed to inflate. From the outflow relation (27),

$$f_{\text{out}} = f_{\text{in}} - \tau_0\, \frac{dv}{dt} = 1$$

Therefore, the outflow also remains the same as that at rest. Only oxygen consumption changes. The deoxyhemoglobin balance (31) becomes

$$\tau_0 \frac{dq}{dt} = m(t) - q$$

Writing $q = 1 + \delta q$ and $m = 1 + \delta m$,

$$\tau_0 \frac{d\,\delta q}{dt} + \delta q = \delta m(t) \tag{38}$$

This linear first-order equation is soluble. Writing $g_0 \equiv 1/\tau_0$ for the washout rate constant,

$$\frac{d\,\delta q}{dt} + g_0\,\delta q = \frac{\delta m(t)}{\tau_0}$$

This gives

$$\delta q(t) = \frac{1}{\tau_0}\int_0^t e^{-(t-t')/\tau_0}\ \delta m(t')\,dt' \tag{39}$$

This is a convolution of the metabolic time course with the kernel

$$K(t) = \frac{1}{\tau_0}\,e^{-t/\tau_0}, \qquad \int_0^\infty K\,(t)\,dt = 1$$

K has a direct physical interpretation: it is the fraction of the deoxyhemoglobin produced at one instant that is still present in the venous compartment a time t later.

In the same delay interval $v = 1$, the signal equation (12) simplifies to

$$\frac{\Delta S}{S} = V_0\left[k_1(1-q) + k_2\left(1 - \frac{q}{v}\right) + k_3(1-v)\right] = V_0(k_1 + k_2)(1-q) = -V_0(k_1 + k_2)\,\delta q(t)$$

The volume-weighted term $k_3(1-v)$ vanishes because the balloon has not inflated. The BOLD signal in this window is therefore proportional to δq and negative.

Substituting the metabolic drive (36), $\delta m(t) = (a_m/\tau_m^2)\,t\,e^{-t/\tau_m}$, and writing $\lambda \equiv 1/\tau_m$ for the decay rate of the metabolic response alongside $g_0 = 1/\tau_0$ for the washout rate,

$$\delta q(t) = \frac{a_m\, g_0}{\tau_m^2} \int_0^t e^{-g_0(t-t')}\; t'\, e^{-\lambda t'}\, dt'$$

$$\delta q(t) = \frac{a_m\, g_0}{\tau_m^2}\, e^{-g_0 t} \int_0^t t'\, e^{\mu t'}\, dt', \qquad \mu \equiv g_0 - \lambda$$

By integrating by parts,

$$\int_0^t t'\, e^{\mu t'}\, dt' = \frac{t\, e^{\mu t}}{\mu} - \frac{e^{\mu t} - 1}{\mu^2}$$

$$\delta q(t) = \frac{a_m\, g_0}{\tau_m^2} \left[\frac{t\, e^{-\lambda t}}{\mu} - \frac{e^{-\lambda t} - e^{-g_0 t}}{\mu^2} \right]$$

Substituting $g_0 = 1/\tau_0$ gives equation (40):

$$\delta q(t) = \frac{a_m}{\tau_m^2 \tau_0} \left[\frac{e^{-g_0 t} - e^{-\lambda t}}{(g_0 - \lambda)^2} + \frac{t\, e^{-\lambda t}}{g_0 - \lambda} \right] \tag{40}$$

Within the preflow window $0 \le t \le \Delta$, equation (39) gives the deoxyhemoglobin perturbation as the convolution of the metabolic drive with the washout kernel. Suppose that oxygen consumption rises instantaneously at $t = 0$ to a new constant level and stays there so that $\delta m(t') = \delta m_0$ for all $t' > 0$. This is the idealization in which metabolism is switched on abruptly and flow has not yet responded. Because δm_0 is constant, it comes outside the integral,

$$\delta q(t) = \frac{\delta m_0}{\tau_0} \int_0^t e^{-\frac{t - t'}{\tau_0}}\, \mathrm{d}t' = \delta m_0 \left(1 - e^{-t/\tau_0}\right)$$

Therefore, deoxyhemoglobin accumulates toward a plateau of δm_0 with time τ_0.

The minimum dip is the value reached at the moment the flow arrives, that is, at $t = \Delta$. Applying the delay-window signal relation above,

$$\mathrm{dip}(\Delta) = \left.\frac{\Delta S}{S}\right|_{t=\Delta} = -V_0(k_1 + k_2)\, \delta q(\Delta) = -V_0(k_1 + k_2)\, \delta m_0 \left(1 - e^{-\Delta/\tau_0}\right)$$

For $\Delta \ll \tau_0$

$$\mathrm{dip}(\Delta) \approx -V_0(k_1 + k_2)\,\delta m_0\,\frac{\Delta}{\tau_0} \tag{41}$$

For a given metabolic step, the depth of the dip is set by the ratio of the neurovascular delay to the mean transit time so that the measured dip constrains Δ/τ_0.

Equation 41 assumes an abrupt step in $CMRO_2$. For the smooth drive of eq. 36, $\delta m \approx a_m t/\tau_m^2$ at early times, and eq. 39 gives $\delta q(\Delta) \approx a_m \Delta^2/(2\tau_m^2\tau_0)$ for $\Delta \ll \tau_m, \tau_0$, so that the dip then grows quadratically with the delay, as found numerically below (Table 5). In both cases, the depth varies inversely with τ_0.

(l) An analytic form for the intermediate hemodynamic response function

The revised equations describe a linear time-invariant system and can be solved in closed form. We take a step increase in the $CMRO_2$ at $t = 0$ and allow the flow to follow with a finite rise time but initially no delay. We first derive the transfer function from $CMRO_2$ to BOLD without the metabolic filter (36) that transforms neural activity into a $CMRO_2$ time course or delay Δ. The response to a neural event, which contains both, is given below (eq. 59).

Writing all variables as small perturbations about rest and linearizing (26), (27), (31) and (12),

$$\tau_c\,\frac{d\,\delta v}{dt} + \delta v = \alpha\,\delta f, \qquad \tau_c = \alpha(\tau_0 + \tau_v) \tag{42}$$

$$\delta f_{\text{out}} = \delta f - \tau_0\,\frac{d\,\delta v}{dt} \tag{43}$$

$$\tau_0\,\frac{d\,\delta q}{dt} + \delta q = \delta m - \delta f_{\text{out}} + \delta v \tag{44}$$

$$\frac{\Delta S}{S} = -A\,\delta q + B\,\delta v, \qquad A = V_0(k_1 + k_2), \qquad B = V_0(k_2 - k_3) \tag{45}$$

A is the sensitivity of the signal to deoxyhemoglobin content, and B is its sensitivity to volume. Let s denote the Laplace variable and $M(s)$, $F(s)$, $V(s)$, $Q(s)$, and $F_{\text{out}}(s)$ denote the transforms of δm, δf, δv, δq and δf_{out}, respectively. For this intermediate transfer function, the flow is

taken to follow metabolism through a low-pass filter of rise time τ_n and gain n, where n is the steady-state CBF–$CMRO_2$ coupling ratio, the ratio of the fractional change in flow to the fractional change in oxygen consumption. With these definitions,

$$V(s) = \frac{\alpha F(s)}{1+\tau_c s}, \qquad F_{\text{out}}(s) = F(s)\,\frac{1+\alpha\tau_v s}{1+\tau_c s} \tag{46}$$

$$Q(s) = \frac{M(s)-F_{\text{out}}(s)+V(s)}{1+\tau_0 s}, \qquad F(s) = \frac{n\,M(s)}{1+\tau_n s} \tag{47}$$

The collection terms give the hemodynamic transfer function from $CMRO_2$ to BOLD:

$$H(s) = \frac{-A\,N(s)+B\alpha n\,(1+\tau_0 s)}{(1+\tau_0 s)(1+\tau_n s)(1+\tau_c s)} \tag{48}$$

$$N(s) = (1+\tau_n s)(1+\tau_c s) - n(1-\alpha) - n\alpha\tau_v s \tag{49}$$

For a unit step in $CMRO_2$, $M(s) = 1/s$, and partial fractions give one exponential per root. Indexing the three roots by i with time constants $\tau_i \in \{\tau_0, \tau_n, \tau_c\}$,

$$\Delta S(t) = p_0 + \sum_i R_i\; e^{-t/\tau_i} \tag{50}$$

$$R_i = \frac{-\tau_i^2\,P(-1/\tau_i)}{\prod_{j\neq i}(\tau_i-\tau_j)}, \qquad P(s) = -A\,N(s) + B\alpha n\,(1+\tau_0 s) \tag{51}$$

with $p_0 = P(0)$ and $\sum_i R_i = -p_0$ since the response must start from zero. The HRF proper — the impulse response — is the derivative,

$$h(t) = -\sum_i \frac{R_i}{\tau_i}\,e^{-t/\tau_i} \tag{52}$$

a sum of decaying exponentials. Empirical models of the HRF instead assume a functional form, such as gamma or other basis functions (5, 6); here, the form is derived, and each time constant has a physical meaning.

Three specific results can be derived from the transfer function

First, letting $s \to 0$ in (48),

$$\Delta S(\infty) = n[A(1-\alpha) + B\alpha] - A = M[n(\beta_D - \alpha) - \beta_D] \tag{53}$$

where $M = V_0(k_1 + k_3)$ and $\beta_D = (k_1 + k_2)/(k_1 + k_3)$. This is algebraically identical to the calibrated BOLD relation of Davis et al. (18) obtained as an output of the model rather than as an assumption. The numerical results are given in the Results section.

Second, for $\tau_n > 0$,

$$\left.\frac{d\,\Delta S}{dt}\right|_{0^+} = -\frac{A}{\tau_0} = -\frac{V_0(k_1+k_2)}{\tau_0} < 0 \tag{54}$$

This does not depend on n, α, or τ_v. If CBF takes any time at all to rise — even with zero delay — the BOLD response must begin downward because deoxyhemoglobin is being produced at $t = 0^+$ while venous washout has not yet changed. Neurovascular delay can produce a dip but is not necessary.

Third, if there is a simultaneous step ($\tau_n = 0$), the initial slope is

$$\left.\frac{d\,\Delta S}{dt}\right|_{0^+} = \frac{\alpha[\tau_0(Bn-A)+A\tau_v(n-1)]}{\tau_0\,\tau_c} \tag{55}$$

which is also negative — that is, a dip occurs without a delay or a flow rise time — if

$$n < n_{\text{dip}} \equiv \frac{A(\tau_0+\tau_v)}{\tau_0 B+A\tau_v} = \frac{\tau_0+\tau_v}{\tau_v+(B/A)\,\tau_0} \tag{56}$$

The interpretation is that the balloon must inflate before its outflow can rise, so washout of deoxyhemoglobin necessarily lags production. Combining (56) with the requirement that the plateau (53) be positive, $n > \beta_D/(\beta_D - \alpha)$, gives the window within which a transient dip is followed by a positive response,

$$\frac{\beta_D}{\beta_D-\alpha} < n < n_{\text{dip}}, \qquad \text{satisfied only if} \quad \frac{\tau_v}{\tau_0} < \frac{1}{\alpha} - 1 = 1 + \beta \tag{57}$$

a Relationships between the presence of a dip, venous viscoelasticity and neurovascular coupling. The simplification in (57) follows from an exact identity between the signal constants,

$$\beta_D\left(1-\frac{B}{A}\right)=\frac{k_1+k_2}{k_1+k_3}\cdot\frac{(k_1+k_2)-(k_2-k_3)}{k_1+k_2}=\frac{k_1+k_3}{k_1+k_3}=1 \tag{58}$$

such that $B/A = 1 - 1/\beta_D$ exactly and the calibration exponent cancels: a transient dip followed by a positive response requires that the viscoelastic time constant be shorter than $(1+\beta)$ times the mean transit time.

(m) Complete hemodynamic response to a neural event

Above, we described the effect of a step in oxygen consumption on the signal. The expression of the metabolic filter that relates neural impulses to changes in metabolism and the effects of delayed flow are both linear, so they can be combined with the intermediate transfer function, and an expression for the complete chain, from a neural impulse to a BOLD signal, can be derived. With $M(s) = a_m/(1+\tau_m s)^2$ for the metabolic branch and $F(s) = \varepsilon\, e^{-\Delta s}/(s^2 + \kappa s + \gamma)$ for the delayed flow branch, substitution into (46) and (47) gives

$$H(s)=\frac{-A\,a_m}{(1+\tau_m s)^2\,(1+\tau_0 s)}+\frac{\varepsilon\, e^{-\Delta s}[B\alpha(1+\tau_0 s)+A(1-\alpha+\alpha\tau_v s)]}{(s^2+\kappa s+\gamma)(1+\tau_c s)(1+\tau_0 s)} \tag{59}$$

The first term is the metabolic branch acting alone; it has a double root at $s = -1/\tau_m$ and a simple pole at $s = -1/\tau_0$ and inverts to a combination of $t\,e^{-t/\tau_m}$, e^{-t/τ_m} and e^{-t/τ_0}. The second term carries the factor $e^{-\Delta s}$, which is a time shift, so it contributes nothing to $t < \Delta$. The initial dip is therefore the first term acting alone.

The flow term depends on the roots of $s^2 + \kappa s + \gamma$. For practical values, these are complex ($\kappa^2 < 4\gamma$), so the flow term is a damped oscillation rather than a simple decay. For the parameters fitted below, the damping ratio $\kappa/(2\gamma^{1/2})$ is 0.6 to 0.9, so the oscillation is damped, and any undershoot it produces in the flow itself is modest. The impulse response is thus

$$h(t) = h_m(t) + h_f(t-\Delta)\,\theta(t-\Delta) \tag{60}$$

with θ the unit step. Both parts can be written out explicitly. Inverting the first term of (59),

$$h_m(t) = -A\,a_m\left[\frac{\tau_0\left(e^{-t/\tau_0}-e^{-t/\tau_m}\right)}{(\tau_m-\tau_0)^2}+\frac{t\,e^{-t/\tau_m}}{\tau_m(\tau_m-\tau_0)}\right] \tag{61}$$

The second term has roots at $s=-1/\tau_0$, $s=-1/\tau_c$ and the conjugate pair $s=-\sigma\pm i\omega_d$, where $\sigma=\kappa/2$ and $\omega_d=\sqrt{\gamma-\kappa^2/4}$. Its inverse is two exponentials and a damped cosine of amplitude R_d and phase ϕ,

$$h_f(t) = R_0\,e^{-t/\tau_0}+R_c\,e^{-t/\tau_c}+R_d\,e^{-\sigma t}\cos(\omega_d t+\phi) \tag{62}$$

Writing $c_0=B\alpha+A(1-\alpha)$, $c_1=\alpha(B\tau_0+A\tau_v)$ and $P=\varepsilon/(\tau_c\tau_0)$, the residues are

$$R_0=\frac{P(c_0-c_1/\tau_0)}{(\tau_0^{-2}-\kappa\tau_0^{-1}+\gamma)(\tau_c^{-1}-\tau_0^{-1})},\qquad R_c=\frac{P(c_0-c_1/\tau_c)}{(\tau_c^{-2}-\kappa\tau_c^{-1}+\gamma)(\tau_0^{-1}-\tau_c^{-1})} \tag{63}$$

$$R_p=\frac{P(c_0+c_1 p)}{2i\omega_d(p+\tau_c^{-1})(p+\tau_0^{-1})},\qquad p=-\sigma+i\omega_d,\qquad R_d=2|R_p|,\qquad \phi=\arg R_p \tag{64}$$

Equations (60)–(64) are the complete analytic hemodynamic response functions of the model, which are the sum of three exponentials and one damped cosine, with the flow contribution switched on at $t=\Delta$. Evaluated for the parameters of Table 4, they give

$$h_m(t) = -2.176\left(e^{-t/1.699}-e^{-t/2.985}\right)-0.552\,t\,e^{-t/2.985} \tag{65}$$

$$h_f(t) = 17.803\,e^{-t/1.699}-1.699\,e^{-t/3.819}+21.052\,e^{-0.4459t}\cos(0.2105\,t-2.442) \tag{66}$$

in units where V_0 is factored out of A and B, with $\varepsilon=0.50$ as set by the peak-CBF constraint. The damped cosine in (66) has a period $2\pi/\omega_d$ of 29.8 s but decays with a time constant $1/\sigma$ of 2.2 s, so it does not appear as ringing. Its negative half-cycle falls within the first 4 s, where it offsets the large positive washout term, and it is positive during the undershoot; the undershoot is instead produced mainly by the negative balloon-relaxation term in τ_c and by the tail of the metabolic branch (see Fig. 7). Plots of the HRF are provided below.

Numerical and analytical results

The theoretical derivations provide analytic expressions for the HRF and allow the HRF to be derived numerically and fitted to the canonical HRF. The results below are all numerical: the revised model equations were first integrated and fitted as described in the *Materials and Methods*. No new experimental measurements are reported; the target for fitting is the canonical HRF used in standard fMRI analysis software, and comparisons are made with physiological values from the literature. The analytic expressions are also illustrated for different conditions and choices of physiological parameters and are shown to be consistent with the other results.

(i) Fitting the revised equations to the canonical HRF

Equations (25)–(28) with the signal equation (12) were fitted to the SPM canonical HRF at 3 T (Table 1, Fig. 1). Two physiological constraints were imposed as penalties on the objective function so that the fitted parameter set reproduces the canonical shape while also operating at a realistic response amplitude with physically admissible vessel mechanics.

First, the peak increase in CBF is held at +40%. Second, the venous stiffness index is required to satisfy $\beta \geq 0$ since a negative β represents a vessel whose pressure falls as it inflates. With a signal constant of (13) at 3 T and a resting venous blood volume fraction $V_0 = 0.03$, a +40% peak CBF increase produces a peak BOLD change of 0.63%, which lies in the 0.5–2% range typical for evoked responses at this field strength and echo time. Left unconstrained, the fit would drive the neural efficacy ε toward zero and return a parameter set that describes the curve well but corresponds to a vanishingly small physiological response.

Table 1. Parameters of the constrained fit to the canonical HRF.

Parameter	Symbol	Value	Interpretation
Vasoactive signal decay	$\tau_z = 1/\kappa$	1.55	decay of z (s)
Autoregulatory feedback	$\tau_f = 1/\gamma$	3.43	CBF feedback (s²)
Resting mean transit time	τ_0	4.55	venous MTT (s)
Venous stiffness index	β	0.00 (at bound)	dimensionless
Inflation viscosity	τ_{in}	0.00	s
Deflation viscosity	τ_{de}	7.52	s; drives the undershoot
Neural efficacy	ε	0.43	sets response amplitude
Resting extraction	E_0	0.34 (fixed)	OEF at rest
Implied Grubb exponent	$\alpha = 1/(2+\beta)$	0.500	derived, not fitted
Fit quality	R^2	0.9945	vs canonical HRF

A fitted α of 0.50 lies above Grubb's 0.38, and a fitted τ_0 of approximately 4.5 s exceeds the 2 s conventionally assumed.

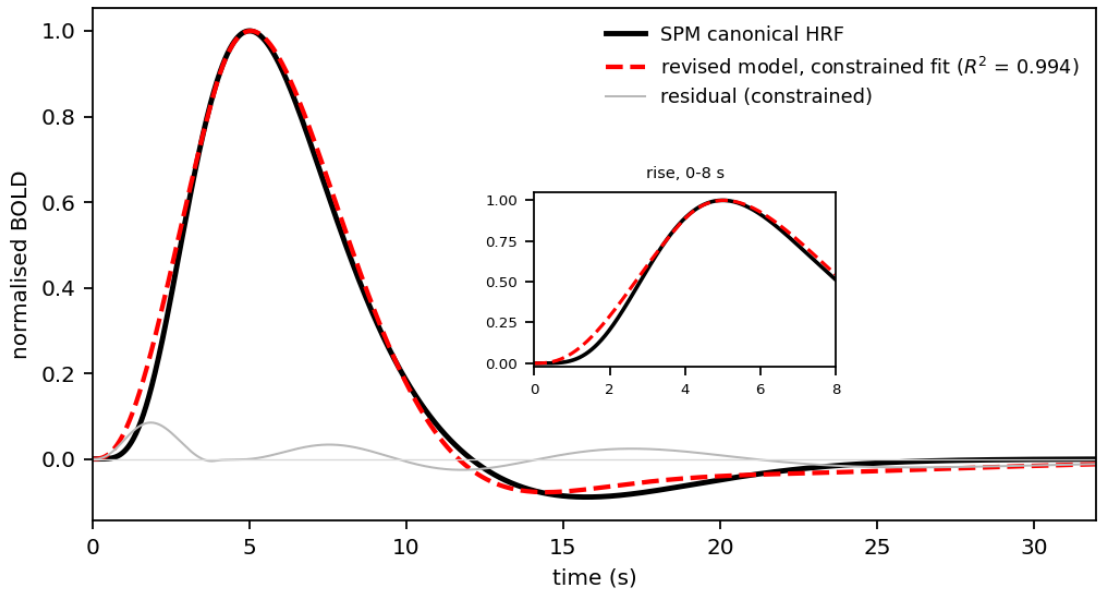


Figure 1. Fit of the revised model to the canonical HRF. *The fitted curve (red, dashed) is the constrained solution of Table 1, with the peak CBF held at +40% and* $\beta \geq 0$, $R^2 = 0.994$. *The light gray trace is the residual, whose largest feature is a small early overshoot at 1–2 s. Inset: the rising phase, 0–8 s.*

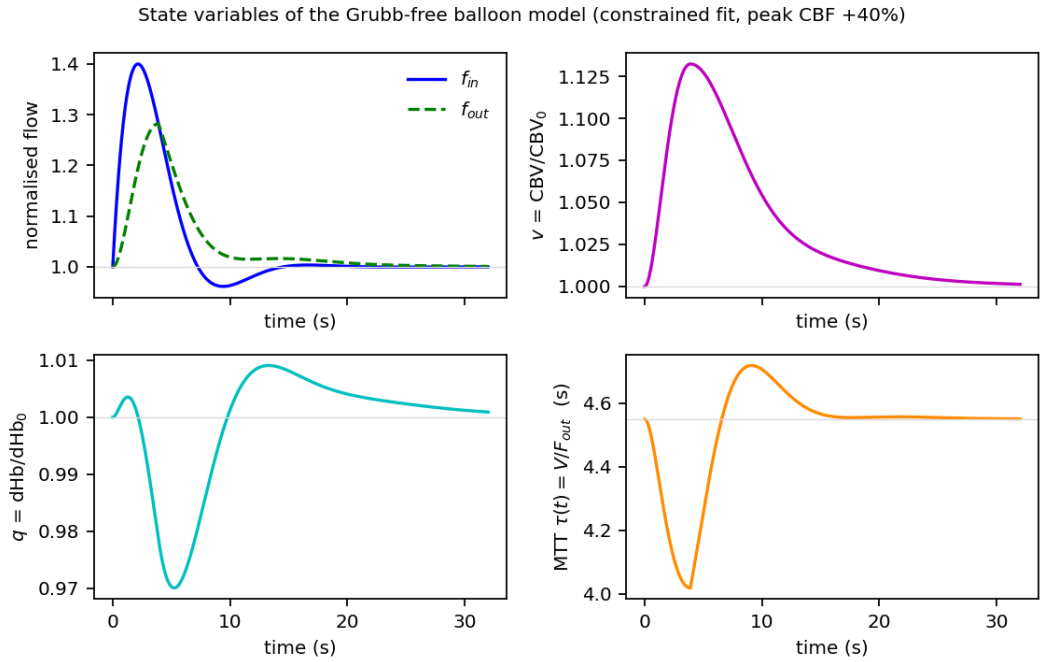
Table 2 lists the physiological quantities predicted at this operating point, and Fig. 2 shows their time courses during the hemodynamic response. The venous volume peaks 1.8 s after flow and

returns far more slowly; the mean transit time falls from 4.55 s to 4.02 s and then overshoots. The +13% CBV response to a +40% CBF response corresponds to an effective α of approximately 0.37 measured dynamically, below the static value of 0.50, because the balloon never reaches a steady state during a brief event.

Table 2. Predicted physiology at the constrained operating point.

Quantity	Value	Comment
Peak CBF	+40%	imposed constraint
Peak CBV	+13.2%	lags CBF by 1.8 s
Minimum total deoxyhemoglobin	−3.0%	at 5.2 s
MTT at peak flow	4.02 s (−12%)	from 4.55 s at rest
Peak BOLD ($V_0 = 0.03$)	0.63%	plausible for 3 T event-related
Volume relaxation$\alpha(\tau_0 + \tau_{\mathrm{de}})$	6.04 s	sets undershoot duration

The model is nonlinear in drive amplitude: the relative undershoot shrinks from −0.13 to −0.02 as the drive increases sixteenfold (Fig. S1), which is one reason why a single canonical HRF cannot be corrected across stimulus intensities (7, 23).



549

550 ***Fig. 2. Physiological variables at the constrained operating points.*** *(Top left) Normalized inflow and*
551 *outflow; (top right) venous volume; (bottom left) total deoxyhemoglobin; (bottom right) instantaneous*
552 *mean transit time* $\tau(t) = V/F_{out}$.

553 **Comparison of the analytic expression with the canonical HRF**

554 Equations (61)–(64) give $h(t)$ explicitly. The canonical HRF contains no initial dip, so for
555 comparison, the metabolic gain was set to $a_m = 0$; the response then decreased to the flow
556 branch (62) alone, a sum of two exponentials and a damped cosine.

557 The three-term expression reproduces the canonical response with $R^2 = 0.998$ (Table 3, Fig. 3),
558 the largest residual being 5.8% of the peak and confined to the first two seconds. The fitted peak
559 falls at 4.94 s against the canonical 5.00 s, and the undershoot reaches -0.082 against -0.089.
560 This result is a closer match than the numerically integrated nonlinear model ($R^2 = 0.994$, Table
561 1).

The parameters returned are physiologically reasonable and were obtained independently of those in Table 1 from the linearized solution rather than from the nonlinear integration. The transit time $\tau_0 = 1.97$ s is close to the 2 s conventionally assumed and considerably shorter than the 4.55 s of the integrated fit; the stiffness index again goes to $\beta = 0$, giving $\alpha = 0.50$; and the damped cosine has $\sigma = 0.406$ s^{-1} and $\omega_d = 0.316$ rad s^{-1}, a period of 19.9 s. Because $1/\sigma$ = 2.5 s is much shorter than the period, the oscillation is heavily damped and does not appear as ringing. At the undershoot minimum (15 s), essentially all of the negative signals originate from the balloon relaxation term in τ_c (−0.088 of the peak), with the damped cosine contributing −0.002 and the washout term +0.009; thus, the undershoot here reflects delayed compliance rather than oscillatory flow.

Table 3. Parameters of the analytic HRF (62) fitted to the canonical response.

Parameter	Symbol	Value	Interpretation
Vasoactive signal decay	$\tau_z = 1/\kappa$	1.23	decay of z (s)
Autoregulatory feedback	$\tau_f = 1/\gamma$	3.77	CBF feedback (s²)
Resting mean transit time	τ_0	1.97	venous MTT (s)
Venous stiffness index	β	0.00	dimensionless
Viscoelastic time constant	τ_v	10.10	s; drives the undershoot
Neural efficacy	ε	0.74	sets response amplitude
Neurovascular delay	Δ	1.26	s; onset lag
Metabolic gain	a_m	0 (fixed)	no dip in the canonical HRF
Implied Grubb exponent	$\alpha = 1/(2+\beta)$	0.500	derived, not fitted
Volume relaxation	$\tau_c = \alpha(\tau_0 + \tau_v)$	6.03	derived, not fitted
Damping of the cosine	$\sigma = \kappa/2$	0.406	s^{-1}
Damped frequency	ω_d	0.316	rad s^{-1} (period 19.9 s)
Fit quality	R^2	0.9983	vs canonical HRF

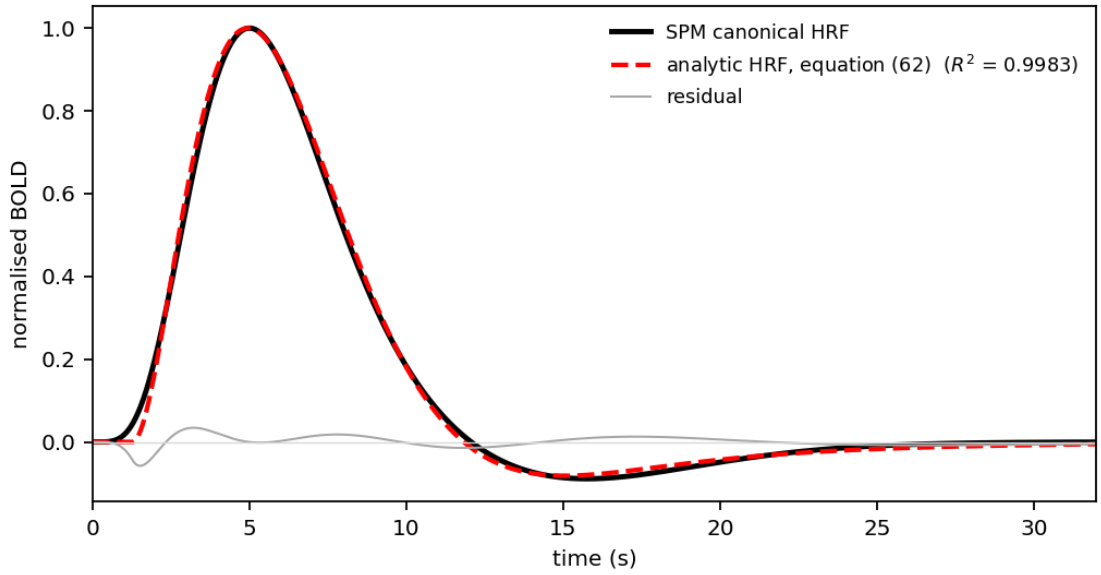


Fig. 3. The analytic HRF was fitted to the canonical response. *The closed form (62), a sum of two exponentials and a damped cosine, is fitted directly to the SPM canonical HRF (black) with the metabolic gain set to zero. The residual (gray) is largest during the first two seconds and nowhere exceeds 5.8% of the peak.*

Using the constrained parameters of Table 1 ($\tau_0 = 4.55$ s, $\tau_v = 7.52$ s, $\alpha = 0.5$; hence, $\tau_c = 6.04$ s, and $n = 2.2$) and a simultaneous step in flow ($\tau_n = 0$), the step response (50) reduces to a two-term exponential,

$$\frac{\Delta S(t)}{\delta m} = 0.03792 - 0.14039\, e^{-t/4.549} + 0.10247\, e^{-t/6.036} \tag{67}$$

The corresponding impulse response changes sign at $t = 11.0$ s, the onset of the poststimulus undershoot. The plateau of (67) is 3.792% per unit relative $CMRO_2$ step, identical to the value returned by the Davis relation (35) with $M = 0.0932$ and $\beta_D = 1.256$, confirming (53).

For the metabolism-first parameters of Table 4 ($\tau_0 = 1.70$ s, $\tau_v = \tau_{de} = 10.03$ s), the zero-delay threshold from (56) is $n_{dip} = 1.13$, which is well below the steady-state coupling $n = 3.80$ (and the peak ratio $n_{peak} = 5.97$); thus, without a neurovascular delay, no dip occurs, consistent with the first row of Table 5.

The impulse responses obtained from (52) are shown in Fig. 4. With a finite flow rise time, the initial slope is $-A/\tau_0 = -2.57\%$ per second per unit $CMRO_2$ step, independent of n, α and τ_v, as required by (54); the resulting dips are −0.52%, −0.98% and −1.74% of the step for rise times of 0.5, 1 and 2 s, respectively, with nadirs at 0.49, 0.92 and 1.66 s, respectively (Fig. 5). For the Table 1 parameters, the two thresholds in (57) are 1.661 and 1.429, so the window is closed, and a simultaneous step cannot produce a transient dip; the window opens once $\tau_v < (1+\beta)\tau_0 = 4.55$ s.

Both analytic solutions were verified against direct numerical integration of the full nonlinear equations. Linearization is reliable for calculating the few percent of $CMRO_2$ changes typical of resting-state fluctuations but should be used with care for strong evoked responses.

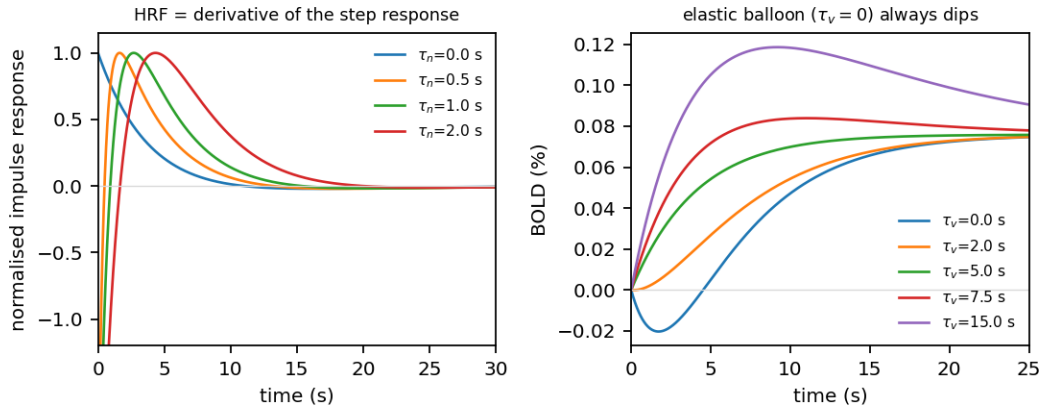


Fig. 4. The analytic HRF for a range of flow rise times. *(Left) Impulse responses from equation (52) for flow rise times τ_n of 0 to 2 s; with any finite rise time, the response becomes negative, as required by equation (54). (Right) Step responses showing that an elastic balloon (τ_v= 0) produces a dip even when the CBF steps simultaneously with the CMRO2 and that the dip is progressively abolished as delayed compliance is added.*

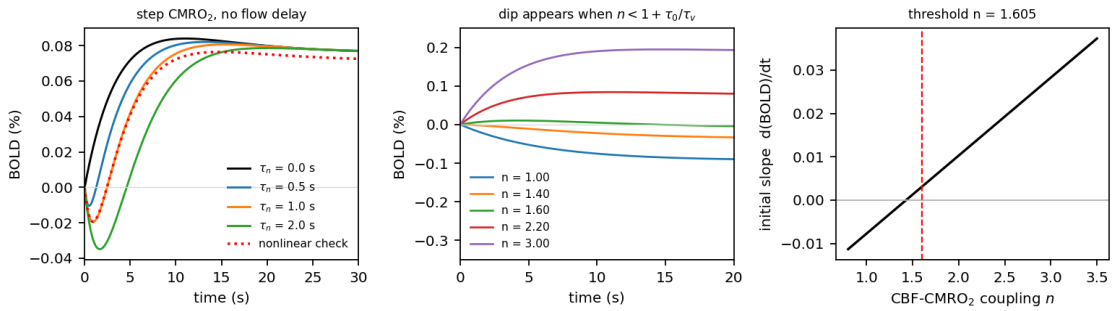
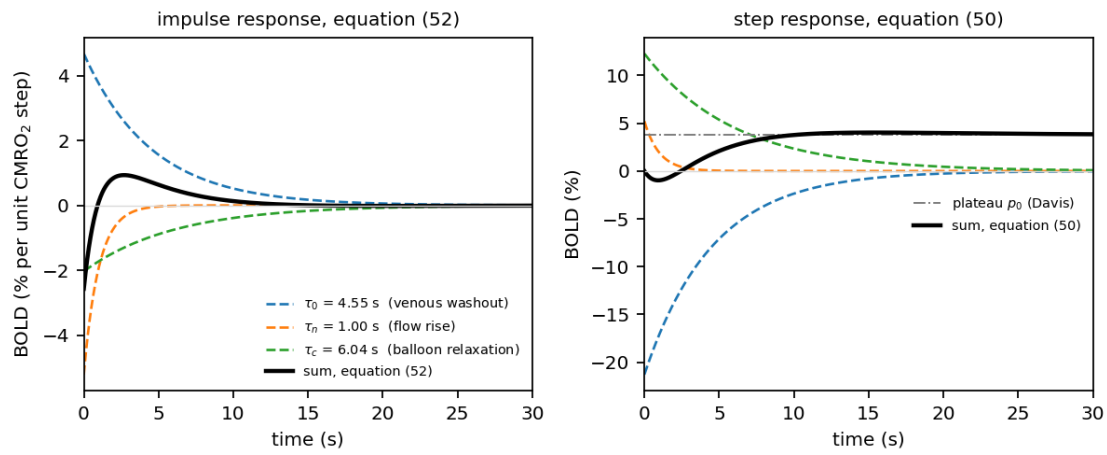


Fig. 5. Validation of the analytic solution and the dip conditions *(Left) Analytic step responses from equations (48)–(51) for flow rise times of 0–2 s, with a numerical integration of the full nonlinear model overlaid (dotted). (Center) Effect of the coupling ratio n. (Right) The initial slope changes sign at* n_{dip} *= 1.429, as predicted by equation (56); the dashed line marks the value 1 +* τ_0/τ_v *obtained if the volume term is neglected.*

Equation (52) expresses the HRF as a sum of three decaying exponentials, one for each time constant of the system, and Fig. 6 plots those terms separately together with their total for a representative flow rise time of $\tau_n = 1$ s. The individual terms are much larger than the response they produce: the largest residue is 5.6 times the plateau. The venous washout term in τ_0 is positive, and the flow and balloon terms are negative; thus, the sign of early response is set such that the patient decays fastest. With a finite flow rise time, the flow term dominates at $t = 0^+$, and the response begins at $-A/\tau_0 = -2.57\%$ per second, as (54) requires that the washout term then reaches a positive peak of 0.93% at 2.69 s, and the slowest term, the balloon relaxation in $\tau_c = 6.04$ s, lasts longer and produces the crossing back through zero at 15.4 s, which becomes the poststimulus undershoot.



621

622 ***Fig. 6. The analytic HRF is the sum of the exponentials.*** *The terms of equation (52) (left, impulse*
623 *response) and equation (50) (right, step response) are plotted separately (dashed) together with their sum*
624 *(black) for the constrained parameters of Table 1 and a flow rise time* τ_n*= 1 s. The colors identify the*
625 *three time constants: venous washout* τ_0*= 4.55 s, flow rise* τ_n *= 1 s and balloon relaxation* τ_c *= 6.04 s.*
626 *The individual terms are several times larger than the response itself and cancel at t = 0; the dashed gray*
627 *line in the right panel is the plateau* p_0*, which equals the Davis value.*

628 The same decomposition applied to the complete response to a neural event is shown in Fig. 7.

629 Equations (65) and (66) are the numerical forms of the two branches at the metabolism-first

630 operating point of Table 4, and each is a sum of three terms: for the metabolic branch, two

631 exponentials and the $t\,e^{-t/\tau_m}$ term; for the flow branch, two exponentials and the damped

632 cosine. The right panel shows how the two branches combine: the metabolic branch acts alone

633 for $t < \Delta$ and is wholly responsible for the dip, -0.063 at 1.23 s; the delayed flow branch then

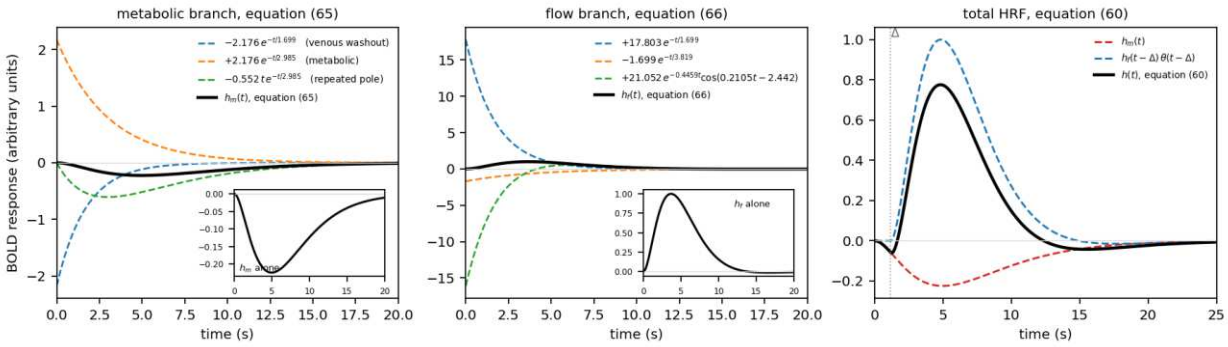
634 dominates, producing a peak of 0.776 at 4.80 s; and the undershoot of -0.042 at 15.5 s arises

635 from the viscoelastic balloon-relaxation term of the flow branch (−0.040) and the slowly

636 decaying tail of the metabolic branch (−0.036), partly offset by the damped cosine (+0.029),

637 which is positive at that time. Each of the three features of the HRF is thus attributable to an

638 identifiable term of the analytic solution.



639

640 ***Fig. 7. The complete HRF is the sum of the exponentials.*** *(Left) The metabolic branch (65): two*
641 *exponentials and the* $t\,e^{-t/\tau_m}$ *term from the repeated pole, with their sum* $h_m(t)$ *in black; the inset*
642 *repeats the sum on its own scale. (Center) The flow branch (66): two exponentials and the damped cosine*
643 *from the complex pole pair, with their sum* $h_f(t)$*; note that the individual terms reach 17.8, while the sum*
644 *peaks at 1.0. (Right) The two branches combined as in (60); the flow branch delay* $\Delta = 1.15$ *s (dotted*
645 *vertical line). The metabolic branch alone produces the initial dip, the flow branch produces the peak,*
646 *and the balloon-relaxation term of the flow branch and the tail of the metabolic branch produce the*
647 *undershoot.*

648 **(ii) Calibration constants derived from the model**

649 The constants that are normally assumed to calibrate BOLD signals are derived and may be

650 compared with the values conventionally adopted.

651 The derived exponent $\beta_D = (k_1 + k_2)/(k_1 + k_3) = 3.900/3.105 = 1.256$ may be compared

652 with the value of approximately 1.3 conventionally assumed at 3 T in calibrated BOLD (18–20,

653 44).

654 With $\alpha = 0.50$, $\beta_D = 1.256$ and $n = 2.2$, the metabolic gain is $G = 0.407$, and the sign-reversal

655 threshold is $n_{\text{crit}} = 1.66$.

656 **(iii) Fitting an HRF with an initial dip**

657 The metabolism-first formulation (eqs. 26, 27, 31, 36 and 37) was fitted to the canonical HRF

658 with an early negative lobe added, scaled so that the nadir was −12% of the positive peak, under

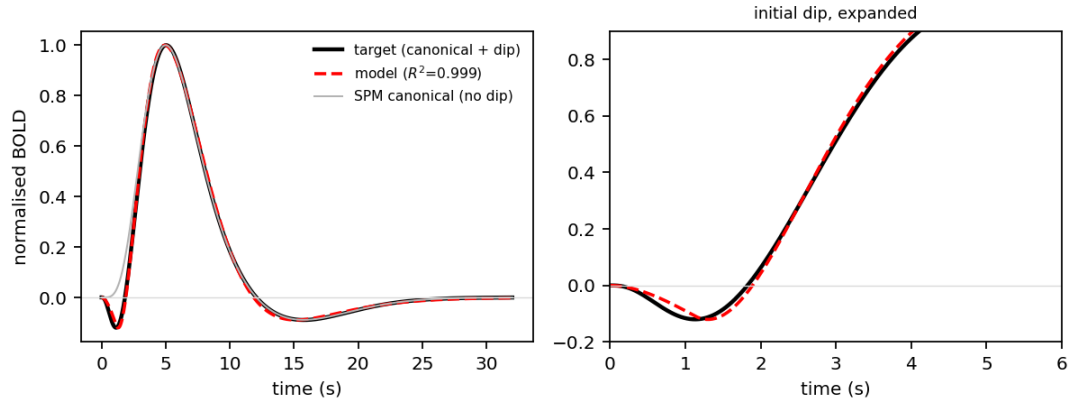
659 the same +40% peak-CBF constraint and with the added physical requirement that OEF never

exceeds unity. All ten parameters were refitted (Table 4; response features in Table S3). The dip nadir occurs at 1.30 s, the positive peak occurs at 4.94 s, and the late undershoot occurs at 15.2 s (Figure 8). Oxygen extraction increases from 0.340 to 0.356 during the delay and then falls to 0.260 once the flow arrives (Fig. 9).

Table 4. Parameters for the metabolism-first fit to a canonical HRF with a 12% initial dip.

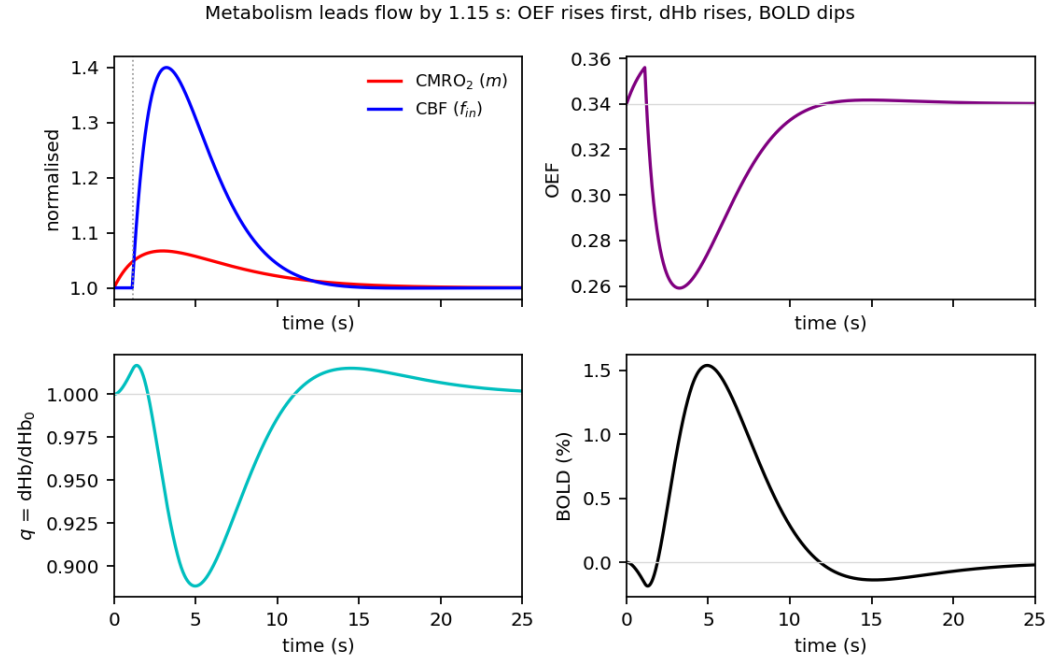
Parameter	Symbol	Value	Comment
Flow onset delay	Δ	1.15 s	metabolism leads flow
Metabolic time constant	τ_m	2.99 s	$CMRO_2$ rise
Resting mean transit time	τ_0	1.70 s	4.55 s in Table 1
Venous stiffness index	β	1.07	0.00 in Table 1
Implied Grubb exponent	α	0.326	0.500 in Table 1
Deflation viscosity	τ_{de}	10.03 s	late undershoot
Inflation viscosity	τ_{in}	0.47 s	—
Vasoactive signal decay	τ_z	1.12 s	flow pathway
Autoregulatory feedback	τ_f	4.11 s²	flow pathway
Metabolic gain	a_m	0.54 s	drive into equation (36)
Coupling ratio (steady state)	n	3.80	ratio of areas, $\int \delta f / \int \delta m$
Coupling ratio (peaks)	n_{peak}	5.97	max δf/maxδm
Fit quality	R^2	0.9992	vs canonical plus 12% dip

Allowing metabolism to increase ahead of flow reduces τ_0 from 4.55 s down to 1.70 s and raises β from 0 to 1.07 so that the implied Grubb exponent falls from 0.50 to 0.326. Both revised values are closer to the independent measurements — a venous MTT of approximately 2 s and a venous-specific α of 0.2–0.4 (30) — than are those in Table 1. In these fits, tying oxygen consumption to flow biased the estimated transit time and flow–volume exponent upward. The two coupling ratios in Table 4 differ because the flow and metabolic responses have different shapes and onset times.



672

673 ***Fig. 8. Fit of the metabolism-first model to a canonical HRF with an initial dip.*** *The right panel shows*
674 *the expansion of the first six seconds.*



675

676 ***Fig. 9. Mechanism of the initial dip.*** *CMRO2 (red) increases 1.15 s before CBF (blue), so oxygen*
677 *extraction must increase to meet demand; total deoxyhemoglobin therefore increases, and the BOLD*
678 *signal becomes negative. When the flow arrives, the extraction coefficient falls to 0.26, and a familiar*
679 *positive response follows.*

680 Holding all the other parameters fixed and varying only the flow delay Δ gives the relations in

681 Table 5 (Fig. S3). The nadir latency tracks the delay almost exactly, $t_{\mathrm{dip}} \approx \Delta + 0.1\mathrm{s}$, so the

682 timing of the dip is a direct readout of the neurovascular delay. The dip depth grows

quadratically at small Δ and closer to linear thereafter, so a measured depth of 10% implies $\Delta \approx$ 1.0 s, and 20% implies $\Delta \approx 1.7$ s. The positive peak is itself delayed approximately one-for-one with Δ at 1.1 s per second, so that fitting a fixed canonical HRF across regions with different neurovascular delays will misestimate both latency and amplitude. This difference may be related to the reduced magnitudes, delayed onset and prolonged initial dips reported in white matter tracts relative to the cortex (16, 41).

Table 5. Dependence of the initial dip on the flow delay

Flow delay Δ (s)	Dip/peak	Nadir latency (s)	Peak latency (s)
0.0	0.000	no dip	—
0.4	−0.022	0.50	4.08
0.8	−0.071	0.94	4.54
1.0	−0.100	1.16	4.78
1.2	−0.129	1.36	5.00
1.6	−0.188	1.76	5.46
2.0	−0.242	2.14	5.90
3.0	−0.340	3.08	6.96

Discussion

We have derived explicit analytic expressions for the HRF of a lumped balloon model of the BOLD effect, requiring only linearization about rest. Three results should be emphasized.

First, Grubb's exponent is not an independent property of the vasculature. Equation (17) shows that the Grubb relationship describes how the mean transit time of blood shortens when flow increases, and (20) shows that within a simple mechanical picture, it is fixed by a single distensibility exponent. Thus, the quantities that actually modulate the hemodynamic response are τ_0, which is directly measurable, and the stiffness of the draining vessels, a property that may differ between gray and white matter, between brain regions, and with age or vascular disease.

Second, the source term in the previous balloon model was the normalized $CMRO_2$, by which oxygen consumption changes in parallel with flow. The original balloon model can still produce an early negative deflection, but this deflection is a vascular transient rather than the metabolic mismatch that the initial dip is considered to signify. Replacing that term with an independently specified term $m(t)$, as in earlier generalizations (21, 23), does not affect steady-state values but changes the transient response, and the analytic solution makes the consequences explicit: with any finite flow rise time, the response must begin downward with a slope $-V_0(k_1 + k_2)/\tau_0$ that is independent of coupling, stiffness and viscoelasticity (54); even with a simultaneous step, a dip occurs whenever $n < n_{\text{dip}}$ (56), simply because the balloon must fill before its outflow can rise. Whether the dip is detectable at conventional field strengths is a separate question (39, 40).

Third, the reformulation returns the calibrated BOLD relations as outputs. Equation (53), derived by taking $s \to 0$ in the transfer function, is algebraically the Davis relation, and $\beta_D = 1.256$ follows from the Obata signal constants alone rather than being assumed.

An immediate application of our revised model is to tissues in which the HRF is known to differ from that of the cortex. BOLD signals are reliably detected in white matter (15, 17), where HRF has a reduced magnitude, delayed onset and a more prominent initial dip and varies in shape along and between tracts (16, 41, 45). Within the present framework, these are not independent observations: a longer neurovascular delay lengthens the dip and delays the peak (Table 5), while a smaller venous blood volume fraction is expected from the lower microvessel density of white matter (46), and a longer transit time reduces the amplitude and slows the response. White matter also has a distinct energy budget (47), and its BOLD responses are modulated by myelination (48), effects that could affect $m(t)$ and n.

The origin of the poststimulus undershoot has not been clarified (31, 32). The present model arises mainly from delayed compliance, the mechanism proposed by Mandeville et al. (33) and developed by Kong et al. (34) and Zheng and Mayhew (35). Blockley et al. (49) reported that explaining their undershoot in this way required blood volume to take more than 55 s to return to baseline after a 4.8 s stimulus, which they considered implausible, and discussed a poststimulus increase in oxygen metabolism as an alternative; poststimulus neuronal activity may also contribute to this change (50). The fits reported here do not require slow volume recovery; after a brief event, the venous volume in the numerical solutions shown in Tables 1 and 4 returns to within 10% of its peak excursion within 17 to 18 s, consistent with linearized volume time constants τ_c of 6.0 and 3.8 s, respectively. Moreover, in the metabolism-first solution, the tail of the metabolic branch, that is, the $CMRO_2$ that remains elevated after flow has returned, contributes −0.036 to the undershoot, comparable to the −0.040 from balloon relaxation (Fig. 7). The two mechanisms are therefore not mutually exclusive.

Limitations

A single compartment represents venous, capillary and arterial volume changes. The arterial CBV responds rapidly, whereas the venous CBV changes more slowly (51), allowing the arterial volume to vary and improving the fit with contrast agent data (49); moreover, layer-specific VASO reveals substantial capillary and arteriolar volume changes (52). Arterial contributions to the total CBV response are thus well established (30, 36). The single compartment is therefore an effective surrogate description appropriate for modeling conventional-resolution gradient-echo BOLD effects but not, in reality, a purely venous mechanical compartment. If compliant upstream volume changes are incorporated, the compartment will appear more compliant than the venous bed, which inflates the fitted τ_0 and α and biases β downward. This may explain why

744 the fits of Tables 1 and 3 reduce β to its lower bound. Three-compartment and
745 multicompartment models (53, 54) could separate these contributions into more complex models.

746 The coupling ratio n is treated as fixed; if it varies with baseline metabolism, as some calibrated
747 BOLD data suggest, then G is not constant, and (35) acquires an additional term. The value $\beta_D =$
748 1.256 inherits the assumptions behind $\Delta\omega_0$, r_0 and ε_r at 3 T and is a consistency check rather
749 than an independent measurement.

750 The metabolic drive (36) is phenomenological. Only its aggregate output enters the model, so the
751 individual relaxations cannot be identified from BOLD data, and τ_m should be read as a lumped
752 constant describing the overall delay between neural activity and oxidative flux rather than the
753 rate of any single biochemical step.

754 Finally, Δ is modeled as a pure transport lag; a distributed lag would fit comparably and is
755 arguably more physiological but has no closed-form dip expression.

756 Materials and Methods

757 **Numerical integration.** Equations (25)–(28), or (26), (27), (31) and (37), with metabolic drive
758 (36), for the metabolism-first formulation, were integrated with a fourth-order Runge–Kutta
759 scheme at a time step of 20 ms over 0–32 s, with the neural drive $u(t)$ a unit impulse at $t = 0$.
760 The BOLD signal was evaluated according to (12). The metabolic drive (36) was integrated in its
761 two-stage form, $\tau_m\, dx/dt = -x + a_m u(t)$ and $\tau_m\, dy/dt = -y + x$, with $m = 1 + y$, which is
762 the standard way of expressing a second-order response as a pair of first-order equations so that
763 it can be advanced by the same solver as the balloon; x and y are internal to the integration and
764 carry no physiological meaning.

Fitting. The target was the canonical HRF used in SPM, the difference between two gamma densities with shape parameters 6 and 16 and a 1:6 amplitude ratio. The model output was normalized to the unit peak and fitted by multistarting Nelder–Mead simplex minimization of the mean squared difference using 18 starting points comprising three literature-based sets and 15 drawn at random from the bounded parameter space; the best solution was subsequently refined on the fine time grid. A quadratic penalty was added to the constrained fits to enforce a peak CBF increase of +40%, and where stated, a penalty was applied to enforce OEF ≤ 1. Identifiability was assessed by profiling the cost against β and τ_0, with all remaining parameters refitted at each point.

Fixed constants. MR constants were fixed at 3 T values: TE = 30 ms, $\Delta\omega_0 = 80.6$ s^{-1}, $r_0 = 25$ s^{-1}, $\varepsilon_r = 1.43$ and $V_0 = 0.03$, giving $k_1 = 3.535$, $k_2 = 0.365$ and $k_3 = -0.430$. The reference physiological values are listed in Table S2. Resting extraction was fixed at $E_0 = 0.34$ since amplitude normalization leaves it only weakly identifiable from the response shape.

Code availability. All analyses were performed in Python using only NumPy and Matplotlib; the Runge–Kutta integrator and Nelder–Mead optimizers were implemented directly. The code and numerical outputs are provided in *data file S1*.

Supplementary Materials

This PDF file includes:

Figs. S1 to S3

Tables S1 to S4

The other Supplementary Materials for this manuscript include the following:

Data file S1 (code and numerical output)

References and Notes

1. R. B. Buxton, E. C. Wong, L. R. Frank, Dynamics of blood flow and oxygenation changes during brain activation: The balloon model. *Magn. Reson. Med.* **39**, 855–864 (1998).

2. S. Ogawa, T. M. Lee, A. R. Kay, D. W. Tank, Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9868–9872 (1990).

3. B. Biswal, F. Z. Yetkin, V. M. Haughton, J. S. Hyde, Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541 (1995).

4. K. J. Friston, A. Mechelli, R. Turner, C. J. Price, Nonlinear responses in fMRI: The balloon model, Volterra kernels, and other hemodynamics. *NeuroImage* **12**, 466–477 (2000).

5. G. H. Glover, Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage* **9**, 416–429 (1999).

6. M. A. Lindquist, J. M. Loh, L. Y. Atlas, T. D. Wager, Modeling the hemodynamic response function in fMRI: Efficiency, bias and mis-modeling. *NeuroImage* **45**, S187–S198 (2009).

7. D. A. Handwerker, J. M. Ollinger, M. D'Esposito, Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage* **21**, 1639–1651 (2004).

8. P. A. Robinson, et al., BOLD responses to stimuli: Dependence on frequency, stimulus form, amplitude, and repetition rate. *NeuroImage* **31**, 585–599 (2006).

9. P. M. Drysdale, J. P. Huber, P. A. Robinson, K. M. Aquino, Spatiotemporal BOLD dynamics from a poroelastic hemodynamic model. *J. Theor. Biol.* **265**, 524–534 (2010).

10. K. M. Aquino, M. M. Schira, P. A. Robinson, P. M. Drysdale, M. Breakspear, Hemodynamic traveling waves in human visual cortex. *PLOS Comput. Biol.* **8**, e1002435 (2012).

11. K. M. Aquino, P. A. Robinson, M. M. Schira, M. Breakspear, Deconvolution of neural dynamics from fMRI data using a spatiotemporal hemodynamic response function. *NeuroImage* **94**, 203–215 (2014).

12. J. C. Pang, P. A. Robinson, K. M. Aquino, Response-mode decomposition of spatiotemporal hemodynamics. *J. R. Soc. Interface* **13**, 20160253 (2016).

13. N. K. Logothetis, J. Pauls, M. Augath, T. Trinath, A. Oeltermann, Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**, 150–157 (2001).

14. E. M. C. Hillman, Coupling mechanism and significance of the BOLD signal: A status report. *Annu. Rev. Neurosci.* **37**, 161–181 (2014).

15. Z. Ding, et al., Detection of synchronous brain activity in white matter tracts at rest and under functional loading. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 595–600 (2018).

16. M. Li, A. T. Newton, A. W. Anderson, Z. Ding, J. C. Gore, Characterization of the hemodynamic response function in white matter tracts for event-related fMRI. *Nat. Commun.* **10**, 1140 (2019).

17. J. C. Gore, et al., Functional MRI and resting state connectivity in white matter — a mini-review. *Magn. Reson. Imaging.* **63**, 1–11 (2019).

18. T. L. Davis, K. K. Kwong, R. M. Weisskoff, B. R. Rosen, Calibrated functional MRI: Mapping the dynamics of oxidative metabolism. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1834–1839 (1998).

19. R. D. Hoge, et al., Linear coupling between cerebral blood flow and oxygen consumption in activated human cortex. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9403–9408 (1999).

20. N. P. Blockley, V. E. M. Griffeth, A. B. Simon, R. B. Buxton, A review of calibrated blood oxygenation level-dependent (BOLD) methods for the measurement of task-induced changes in brain oxygen metabolism. *NMR Biomed* **26**, 987–1003 (2013).

21. T. Obata, et al., Discrepancies between BOLD and flow dynamics in primary and supplementary motor areas: Application of the balloon model to the interpretation of BOLD transients. *NeuroImage* **21**, 144–153 (2004).

22. R. L. Grubb, M. E. Raichle, J. O. Eichling, M. M. Ter-Pogossian, The effects of changes in $PaCO_2$ on cerebral blood volume, blood flow, and vascular mean transit time. *Stroke* **5**, 630–639 (1974).

23. R. B. Buxton, K. Uludağ, D. J. Dubowitz, T. T. Liu, Modeling the hemodynamic response to brain activation. *NeuroImage* **23**, S220–S233 (2004).

24. K. E. Stephan, N. Weiskopf, P. M. Drysdale, P. A. Robinson, K. J. Friston, Comparing hemodynamic models with DCM. *NeuroImage* **38**, 387–401 (2007).

25. M. Havlicek, et al., Physiologically informed dynamic causal modeling of fMRI data. *NeuroImage* **122**, 355–372 (2015).

26. S. Ogawa, et al., Functional brain mapping by blood oxygenation level-dependent contrast magnetic resonance imaging: A comparison of signal characteristics with a biophysical model. *Biophys. J* **64**, 803–812 (1993).

27. D. A. Yablonskiy, E. M. Haacke, Theory of NMR signal behavior in magnetically inhomogeneous tissues: The static dephasing regime. *Magn. Reson. Med.* **32**, 749–763 (1994).

28. J. L. Boxerman, et al., The intravascular contribution to fMRI signal change: Monte Carlo modeling and diffusion-weighted studies in vivo. *Magn. Reson. Med.* **34**, 4–10 (1995).

29. K. Uludağ, B. Müller-Bierl, K. Uğurbil, An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *NeuroImage* **48**, 150–165 (2009).

30. J. J. Chen, G. B. Pike, BOLD-specific cerebral blood volume and blood flow changes during neuronal activation in humans. *NMR Biomed* **22**, 1054–1062 (2009).

31. P. C. M. van Zijl, J. Hua, H. Lu, The BOLD poststimulus undershoot, one of the most debated issues in fMRI. *NeuroImage* **62**, 1092–1102 (2012).

32. J. Hua, R. D. Stevens, A. J. Huang, J. J. Pekar, P. C. M. van Zijl, Physiological origin for the BOLD poststimulus undershoot in human brain: Vascular compliance versus oxygen metabolism. *J Cereb. Blood. Flow. Metab.* **31**, 1599–1611 (2011).

33. J. B. Mandeville, et al., Evidence of a cerebrovascular postarteriole windkessel with delayed compliance. *J Cereb. Blood. Flow. Metab.* **19**, 679–689 (1999).

34. Y. Kong, et al., A model of the dynamic relationship between blood flow and volume changes during brain activation. *J Cereb. Blood. Flow. Metab.* **24**, 1382–1392 (2004).

35. Y. Zheng, J. Mayhew, A time-invariant visco-elastic windkessel model relating blood flow and blood volume. *NeuroImage* **47**, 1371–1380 (2009).

36. D. A. Boas, S. R. Jones, A. Devor, T. J. Huppert, A. M. Dale, A vascular anatomical network model of the spatiotemporal response to brain activation. *NeuroImage* **40**, 1116–1129 (2008).

37. D. Malonek, A. Grinvald, Interactions between electrical activity and cortical microcirculation revealed by imaging spectroscopy: Implications for functional brain mapping. *Science* **272**, 551–554 (1996).

38. R. B. Buxton, The elusive initial dip. *NeuroImage* **13**, 953–958 (2001).

39. K. Uludağ, To dip or not to dip: Reconciling optical imaging and fMRI data. *Proc. Natl. Acad. Sci. U.S.A.* **107**, E23 (2010).

40. X. Hu, E. Yacoub, The story of the initial dip in fMRI. *NeuroImage* **62**, 1103–1108 (2012).

41. K. G. Schilling, et al., Anomalous and heterogeneous characteristics of the BOLD hemodynamic response function in white matter. *Cereb. Cortex Commun.* **3**, tgac035 (2022).

42. A. B. Simon, R. B. Buxton, Understanding the dynamic relationship between cerebral blood flow and the BOLD signal: Implications for quantitative functional MRI. *NeuroImage* **116**, 158–167 (2015).

43. L. Tang, M. J. Avison, J. C. Gore, Nonlinear blood oxygen level-dependent responses for transient activations and deactivations in V1—Insights into the hemodynamic response function with the balloon model. *Magn. Reson. Imaging* **27**, 449–459 (2009).

44. V. E. M. Griffeth, R. B. Buxton, A theoretical framework for estimating cerebral oxygen metabolism changes using the calibrated-BOLD method. *NeuroImage* **58**, 198–212 (2011).

45. T. Wang, D. M. Wilkes, M. Li, X. Wu, J. C. Gore, Z. Ding, Hemodynamic response function in brain white matter in a resting state. *Cereb. Cortex Commun.* **1**, tgaa056 (2020).

46. T. Kubíková, P. Kochová, P. Tomášek, K. Witter, Z. Tonar, Numerical and length densities of microvessels in the human brain: Correlation with preferential orientation of microvessels in the cerebral cortex, subcortical gray matter and white matter, pons and cerebellum. *J. Chem. Neuroanat.* **88**, 22–32 (2018).

47. J. J. Harris, D. Attwell, The energetics of CNS white matter. *J. Neurosci.* **32**, 356–371 (2012).

48. Y. Gao, L. Xu, K. G. Schilling, et al., Myelination selectively modulates BOLD signal in white matter. Res Sq [Preprint]. rs.3.rs-6597153. doi: 10.21203/rs.3.rs-6597153/v1 (2025)

49. N. P. Blockley, S. T. Francis, P. A. Gowland, Perturbation of the BOLD response by a contrast agent and interpretation through a modified balloon model. *NeuroImage* **48**, 84–93 (2009).

50. K. J. Mullinger, S. D. Mayhew, A. P. Bagshaw, R. Bowtell, S. T. Francis, Poststimulus undershoots in cerebral blood flow and BOLD fMRI responses are modulated by poststimulus neuronal activity. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 13636–13641 (2013).

51. T. Kim, S.-G. Kim, Temporal dynamics and spatial specificity of arterial and venous blood volume changes during visual stimulation: Implication for BOLD quantification. *J. Cereb. Blood Flow Metab.* **31**, 1211–1222 (2011).

52. L. Huber, et al., Validating layer-specific VASO across species. *NeuroImage* **237**, 118195 (2021).

53. Y. Zheng, D. Johnston, J. Berwick, D. Chen, S. Billings, J. Mayhew, A three-compartment model of the hemodynamic response and oxygen delivery to brain. *NeuroImage* **28**, 925–939 (2005).

54. T. J. Huppert, M. S. Allen, H. Benav, P. B. Jones, D. A. Boas, A multicompartment vascular model for inferring baseline and functional changes in cerebral oxygen metabolism and arterial dilation. *J. Cereb. Blood Flow Metab.* **27**, 1262–1279 (2007).

Acknowledgments

Funding: This work was supported by National Institutes of Health grants R01NS129855 (Z.D.), R01NS113832 and R01NS078680 (J.C.G.).

Author contributions:

Conceptualization: J.C.G. and Y.G. Methodology: J.C.G. and Y.G. Formal analysis: T.-W.L. Funding acquisition: Z.D. and J.C.G. Writing—original draft: J.C.G. Writing—review and editing: Y.J., A.L.K., W.W.B., M.L., K.S., A.W.A., and Z.D. Y.G.

Competing interests: The authors declare that they have no competing interests.

Availability of data and materials: All the data needed to evaluate the conclusions of the paper are presented in the paper and/or the Supplementary Materials. All analysis code, the parameter files it writes and the raw numerical output are provided as data file S1.