Per-Pixel Uncertainty and Physical Understanding as the Basis for Trust in Multi-Instrument Products

A white paper submitted to the National Academies ESAS 2028 Request for Information — a contextual topic and implementation approach not specific to any Sphere, developed with snow water equivalent and soil moisture as worked examples

Primary author

Xiaolan Xu — Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA · xiaolan.xu@jpl.nasa.gov

Co-authors

Nicolas Rouquette — Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Amy Braverman — Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Synopsis

Submission classification. This is a **contextual-topic and implementation-approach** submission; it is not a mission or project concept. **Sphere alignment:** not applicable — the topic is not specific to any Sphere. It addresses the request's contextual topic *"approaches for assessing whether third-party data meet science and applications requirements"* and applies to any geophysical quantity that is estimated from more than one instrument. Every instrument measures a physical signal — brightness temperature, backscatter, reflected sunlight, a weight — and the geophysical quantity is derived from it through the measurement physics. Two worked examples are used. Snow water equivalent shows the user side: several measurement principles — radar phase, reflected sunlight, lidar height, pillow weight — are all in operational use, and their products visibly disagree. Soil moisture shows the producer side: a decade of passive microwave data, active radar now in orbit, dozens of physics-based products, and machine-learning products trained on top of them. Nothing in the method depends on the quantity: the same situation holds for aerosol optical depth from several sensors, sea-ice thickness from altimetry and passive microwave, biomass from lidar and radar, or sea-surface temperature from infrared and microwave.

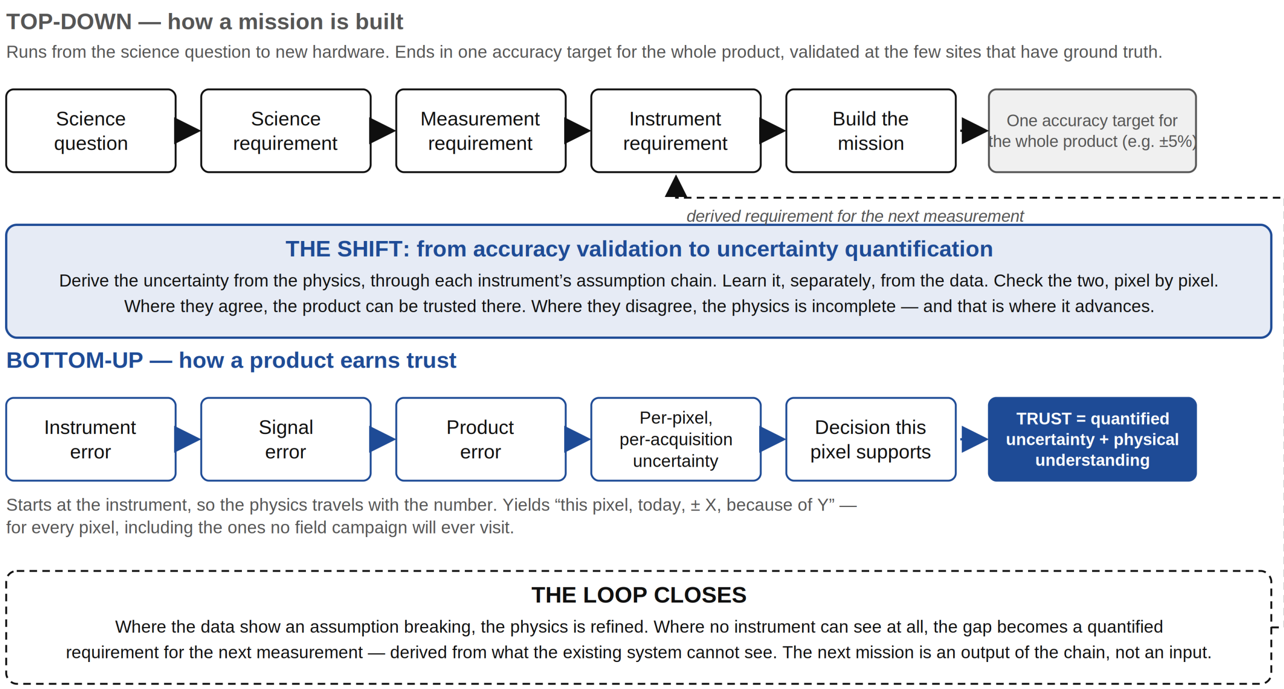

Figure 1 — Traceability in reverse. The top-down chain builds missions and ends in one accuracy target for the whole product. The bottom-up chain starts at the instrument and ends in a per-pixel, per-acquisition uncertainty that carries its physical cause — the two things trust requires.

Trust needs two things. Today's products carry one each.

Soil moisture: a decade of passive microwave, active radar in orbit, dozens of physics-based products, and machine-learning products trained on top of them.

	Physics-based product	Machine-learning product
Physical understanding	YES **Has it.** A measurement principle, stated as assumptions before launch.	NO **Does not have it.** A learned pattern. Nothing says why the number is what it is.
Quantified uncertainty	NO **Not beyond validation.** One accuracy figure, computed only at the sites with ground stations, inside the assumptions. Silent everywhere else.	YES **Has it — statistically.** Learns from all the data at once; sometimes predicts better, and can put a number on its own spread.
Where it breaks	? **Known in principle, never tested.** The assumption says where it should fail. The data that could check it were never used to.	? **Unknown.** No physics behind the pattern, so no way to say where it stops holding — adding physics as a constraint does not change that.

TRUST = the physics that says where a measurement should hold + the quantified uncertainty that says where it actually does

Where the learned uncertainty disagrees with what the physics predicts is exactly where the physics is incomplete — and can be advanced.
That is what the data now make possible: testing the understanding, pixel by pixel, and feeding what breaks back into the physics.

Figure 2 — Two halves of trust, one in each kind of product.

Executive Summary

- **The Data-Rich Challenge:** Users must reconcile conflicting estimates from multiple instruments to determine what to trust in a budget-constrained decade.
- **Per-Pixel Uncertainty:** Trust requires quantifying confidence for every pixel, moving beyond aggregate accuracy validation to continuous uncertainty quantification.
- **Dual-Uncertainty Framework:** Comparing physics-derived and data-learned uncertainties identifies where the product is reliable and where the physics must advance.
- **Decision Support:** Delivering uncertainty as a standard data layer enables operational users to make defensible, data-driven decisions.

In the past, when few missions existed, the community reasoned from a science question down to an instrument that would answer it, and then asked NASA to build that instrument. That reasoning answered its era's question — what to build. The data-rich era asks a new question: when several existing instruments estimate a quantity by different measurement physics, and their products disagree, how does a user know which to trust, where? Snow water equivalent — how much water is stored in the snowpack — is the worked example here. It was named a Targeted Observable by the last survey and recommended as a candidate for its competed Earth System Explorer line (NASEM, 2018). The coming decade is budget-constrained; with few new missions, the emphasis shifts to using existing instruments in new ways. Several instruments that observe different aspects of snow already exist. Operational water managers struggle to reconcile the diverse products derived from them, particularly where products conflict in regions inaccessible for ground-truthing.

A remote-sensing product earns trust from two things: a quantified uncertainty for every value it delivers, and the physical understanding of where that uncertainty comes from. Validation confirms that, within the retrieval's assumptions and at the sites where ground truth exists, the product meets its stated accuracy. What it cannot give — as a number, not only as a statement — is a confidence level for each pixel away from those sites, which is where most pixels are and where a user needs trust most. Soil moisture shows the two halves coming apart: physics-based products carry the understanding but not the uncertainty beyond their validated assumptions; machine-learning products carry a statistical uncertainty but not the physics to understand why. The way forward is to hold both: two uncertainties, one derived from the physics and one learned from the data, checked against each other pixel by pixel. Where they agree, the product can be trusted; where they disagree, the physics is incomplete — and that is where it advances. We need to move from accuracy validation to uncertainty quantification.

Every instrument measures a physical signal. This paper runs that chain the other way: from the instrument's error, through the error of the signal, to the uncertainty of the derived quantity — at each pixel and each acquisition — and finally to the decision that pixel can support. That per-pixel, per-acquisition uncertainty is what operational users need — delivered as a data layer — and it is the output of a bottom-up chain the community has not yet built. Closing that gap is a matter of method and we believe it belongs in a decadal survey. We offer this as a framing for discussion: the practices described are ones we are beginning to apply ourselves, and we would welcome the survey's and the community's judgment on where they hold and where they do not. (Figure 1)

1. Trust in a product needs two things: a quantified uncertainty for every pixel, and the physics that says why

Top-down reasoning answers "what to build" — the data-rich era also asks "what to trust, where"

When few missions existed, the reasoning ran in one direction: from a science question, to a science requirement, to a measurement requirement, to an instrument requirement, to building the mission. This is the science traceability matrix, and it is the reason the instruments below exist. Its end point is hardware; it was not designed to answer which of the products those instruments now produce can be trusted at a given place.

ESAS 2028 sets its own frame: cost-capped, largely competed missions (NASEM, 2026). The decade is budget-constrained and, independently, measurement-rich — which is not the same as being rich in the quantity a user needs. What it lacks is a way to set every measurement principle side by side, each with its uncertainty, so a decision-maker can see what to trust. That is a question the top-down chain needs a complement to answer.

From accuracy validation to uncertainty quantification — the next step builds on the missions we have

The emphasis on accuracy has its reason: it is how a mission concept is formed — the physics stated as assumptions, written into each mission's algorithm basis document together with where they hold, and validated against what can be measured on the ground, so that a mission can promise an accuracy and be built to meet it. That discipline put the instruments of this decade in orbit, and this paper builds on it. What has changed is the data. When a product is global and the archive is decades deep, an assumption written down once can now be tested. The data show, pixel by pixel, where each assumption holds and where it is breaking, and that evidence can be fed back to refine the physics. Quantifying uncertainty is how that feedback runs: it turns the accuracy statement into a per-pixel one, names the assumption responsible where it fails, and shows what no existing measurement can see — where another instrument, or a new one, completes the picture. The earlier way answered its decade's question. This one answers the question a measurement-rich decade asks.

From instrument error to the uncertainty of each pixel, with its cause attached

A gridded product is a model output — a retrieval or an assimilation carries its own assumptions, and error propagates through them as it does through the sensor. The model is a link in the chain.

The top-down chain's error budget — "snow water equivalent to 5% or soil moisture unbias rmse below 0.04 cm3/cm3" — is genuine and valuable. But it is an aggregate: one target for the product as a whole. It does not yield "this pixel, today, plus or minus a stated amount, because of a stated cause," which is what operational users need. Trust is those two things — the uncertainty and its physical cause — and the backward chain is how a product gets both: the uncertainty by construction, and the physics carried with it because the chain started at the instrument.

What the joint observing system cannot see falls out of the backward chain as a quantified observability gap: a requirement a future measurement must satisfy, stated in physical units rather than as advocacy. Top-down asserts what a new instrument must be; bottom-up derives it from what the existing system demonstrably cannot see. Nor is this instrument-driven science: the science question does not change; what changes is that we can say whether the observing system can answer it, and where.

2. Users need to know how far to trust a product — that need is the requirement

The user-side example is operational water management in the western United States, chosen because its users have stated their requirements unusually precisely. Reservoir operations, flood-risk management, water-supply forecasts, hydropower planning, and mandated environmental flows all depend on knowing how much water is stored in the snowpack and when it will reach the river.

The operational obstacle today is not a shortage of snow products — it is the opposite. A forecaster can draw snow water equivalent from a ground network, gridded models, airborne lidar, and satellite retrievals, and the values disagree, often substantially, with nothing in them to explain why. Averaging the products does not converge on truth; it averages the disagreement. When they conflict, the forecaster falls back on judgment — decades of basin knowledge, applied precisely where the products go quiet. That judgment is the most reliable instrument in the system, and the only one that cannot be scaled, transferred, or defended in a coordination meeting.

What operators need is not another estimate. It is a way to know how much to trust the estimate they have: a per-pixel, per-acquisition uncertainty, delivered as a data layer alongside the snow water equivalent itself. Accuracy sits differently in the chain: it is established against reference data that is sparse and itself uncertain, over the few places a comparison is possible. A per-pixel uncertainty is computed for every pixel, including the ones no field campaign will ever visit.

3. Today's products carry one half of trust each — and combining them loses both

Uncertainties do not simply add. A low-uncertainty dataset fused with a high-uncertainty dataset does not produce a low-uncertainty product — and an algorithm operating on finished data products has no way to know that, because the information needed to know it was not carried forward.

Soil moisture illustrates this clearly, with passive L-band (Entekhabi et al., 2010) and active radar (Kellogg et al., 2020) missions providing different signals (brightness temperature vs. backscatter) to derive products via physics-based or machine-learning methods (O and Orth, 2021). These products often disagree (Dorigo et al., 2017), presenting the same reconciliation challenge as snow products.

Physics-based products offer understanding but lack per-pixel uncertainty beyond validated assumptions. Conversely, machine-learning products provide statistical uncertainty but lack physical grounding to identify where learned patterns fail. Constraining learning with physics has not yet reliably overcome this boundary problem.

Two uncertainties, checked against each other: where they disagree is where the physics advances

Trust needs both. The physics says where a measurement should hold and why; the quantified uncertainty says where it actually does. Derive the uncertainty from the physics — propagate the instrument error through the assumptions to every pixel — and, separately, estimate it from the data. The second half already has a method: post hoc uncertainty quantification estimates the error of an operational retrieval after the fact, from the observing system's own data, without changing the retrieval (Braverman et al., 2021); it is the statistical counterpart to the physics-derived chain, and it does not require the physics to be right to run. Then compare the two, pixel by pixel. Where they agree, the product can be trusted there. Where the learned uncertainty differs from what the physics predicts, the physics is incomplete, and that disagreement is the most useful result the data can return: it says which assumption to revisit and where. For the operator the same disagreement is actionable immediately, without waiting for the diagnosis: the pixel is delivered carrying a divergence flag — the physics and the data do not agree here, which points to a process the model does not contain — so it can be down-weighted while the science team works out the cause. The same quantity is a warning downstream and a research target upstream. That is how understanding of the system advances — by finding where it is incomplete — and the data now exist to test it everywhere, not only at the sites with ground truth. This is a shift in how the problem is approached, not a new product: from accuracy validation to uncertainty quantification — from validating assumptions once to quantifying, continuously, where they hold. What the tiers below describe is how that understanding is lost when products are combined. (Figure 2)

- Tier 1 — fusion at the data level. Learn a mapping from many products to a target variable. These systems are genuinely useful, and they are the reason operators have anything gridded at all. What they cannot do is state what the fused variable physically represents — so its error bar has no physical cause it can point to.
- Tier 2 — physically informed fusion. Science models inform the fusion; this is where the most sophisticated work in the field sits today, and each input can carry an uncertainty of its own. But the measurement principles still do not survive the fusion step, so we still cannot say what the fused quantity means to someone about to make a decision.
- Tier 3 — fusion at the model level. Fuse the science models — the physics that produced each dataset, uncertainty and all — rather than their outputs. The result retains the measurement principles of its constituents, and can therefore emit a quality flag whose meaning is defensible. This is not a new class of mathematics — it is the estimation frameworks the community already uses, applied across instruments rather than within one: joint state-space Bayesian estimation, or variational retrieval in which several forward models are inverted simultaneously against the calibrated instrument measurements through their respective observation operators (Rodgers, 2000). What is missing is not the machinery but the declared assumption chains that let one instrument's operator be composed with another's.

Against the two-part criterion for trust, there is a fundamental transition from Tier 1, which relies on statistical mapping at the data level without physical grounding, to Tier 3, which achieves physical causality by fusing measurement principles at the model level. Consequently, Tier 1 delivers neither trust criterion, Tier 2 delivers the uncertainty without the physical understanding, and Tier 3 successfully retains both.

Fusing finished products loses the physics — because nothing in the process yet carries it

Tier 1 and Tier 2 are what one builds when each instrument's assumptions are validated once and not carried forward. With nothing carrying the physics through the combination, fusing at the data level has not been a shortcut — it has been the option available. Both are necessary infrastructure; what is missing is meaning, not computing power.

The uncertainty travels downstream — trust at the product is what makes the science built on it defensible

A product is rarely the end of the chain. Soil moisture feeds precipitation studies, land–atmosphere coupling, drought and agricultural monitoring, wildfire risk, climate-model evaluation, and land-surface change detection, and every one of those inherits the product's error. When the product carries no per-pixel uncertainty, the downstream study cannot separate a real signal from an assumption failing: a soil-moisture trend over a region where vegetation changed may be climate, or may be the retrieval's vegetation assumption breaking. Without the uncertainty and its physical cause, the conclusion drawn is not wrong so much as unfounded. The same holds for any derived quantity in any Sphere, which in our view is why the requirement belongs at the product.

4. Every instrument measures a signal and the quantity is derived — the physics-derived uncertainty comes from writing that chain down

For a satellite instrument, the chain is already written — as assumptions in a document, but not as a number in the product

Each measurement principle can be written out as an explicit assumption chain: what physics the instrument exploits, where that approximation holds, and how error propagates from the raw signal to the geophysical quantity. For a single instrument, the community already writes this down. Every NASA mission publishes an Algorithm Theoretical Basis Document explaining how its instrument's signal becomes a geophysical quantity — the physics it relies on, the approximations it makes, and the conditions under which they hold — and it is a required deliverable (O'Neill et al., 2021).

But what the document holds is a set of assumptions stated in prose, and what the product delivers is a value, a mask, a quality flag, and one accuracy statement for the product as a whole. The assumption is written down; it is not turned into a number. There is no per-pixel, per-acquisition uncertainty in the product that a user can read, that two products can be compared on, or that a fusion step can weight its inputs by. That is the step this paper is about: propagating the instrument's error through the documented chain so that the assumption becomes a quantity that travels with the value.

The same holds, one level up, when instruments are combined. Nothing states, as a number, which of one instrument's assumptions conflict with another's, where one instrument's noise source is another's signal, or what the pair together still cannot see. That is why an operator averaging six snow products cannot discover what drives their disagreement, and it is the practice this paper asks the decade to extend.

Reasoning across the combination this way does two things that fusing finished products cannot: it finds complementarity where one instrument's confounder is exactly what another measures (liquid water destroys radar coherence (Guneriussen et al., 2001), and surface wetness is what imaging spectroscopy determines directly (Bohn et al., 2021)); and it surfaces assumption conflicts before they compound into error.

From assumption to number: demonstrated in research, not yet delivered in any product

The assumptions are understood carefully at each step: the lidar team studies its density model; the interferometry team knows where the dry-snow assumption holds and where it fails; and the ground-network team accounts for the precision of the pressure transducer within the snow pillow. What does not yet happen is turning that understanding into a per-pixel number and carrying it down the chain and through the point where products are combined.

That this can be done has been shown in research for essentially every measurement principle — optical spectroscopy through optimal estimation (Rodgers, 2000; Bohn et al., 2021), microwave through Bayesian and Monte-Carlo retrievals of backscatter and phase (Singh et al., 2024; Bonnell et al., 2024) — and, on the statistical side, post hoc methods have quantified an operational retrieval's error from its own data (Braverman et al., 2021). But these remain demonstrations, per instrument, and none is delivered as a standard quantity in the product. So today no two products can be compared on their uncertainty, and no fusion can weight its inputs by it. The gap is not that the physics is unknown or that the methods do not exist; it is that the assumption has not been made into a number that travels with the value, in a form shared across instruments.

Each expert declares their own instrument's chain, and the composition is performed on those declarations — a separate, auditable step. End-to-end pipelines joining multiple flagship missions already run; what they do not yet carry is a per-pixel uncertainty that is physically attributable across the full chain, because nothing in the process yet requires it.

5. Three gaps between integration and trust

1. **The gap in data access:** A backward chain must begin at the instrument level. Propagating instrument error requires access to calibrated, geolocated measurements, not just finished maps. We must ensure secure, prioritized access to these lower-level data, as all subsequent uncertainty quantification depends on it.
2. **The gap in capability:** NASA's commitment to a "Multisource Integrated Observatory" needs a matching investment in a shared "trust" layer. We need a shared assumption chain that integrates measurements—not just institutions—allowing international or commercial data to enter on equal terms. This shared-framework infrastructure must be resourced as a core capability.
3. **The gap in delivery:** User trust must be computed, not just declared. Unlike airborne platforms with discrete operational limits, satellites return values for every pixel, even where physics may fail. Trust requires an honest, computed error bar—delivered as a per-pixel uncertainty—that enables users to determine for themselves when a product is actionable.

6. Summary

We offer five points for the committee's consideration.

1. Trust in a product needs two things: a quantified uncertainty for every pixel and acquisition, and the physical understanding of where it comes from. Products today carry the assumption in a document and one accuracy statement for the whole product; the step is to turn the assumption into a number that travels with the value, in a form shared across instruments. The enterprise already has the vehicle: propagation of instrument error through the forward model, and a per-pixel uncertainty layer in the delivered product, could be described in the Algorithm Theoretical Basis Document and delivered with it — as the assumptions themselves already are.
2. Two uncertainties — one derived from the physics, one estimated from the data — checked against each other, pixel by pixel. Where they agree, the product can be trusted; where they disagree, the physics is incomplete, and that is where it advances.
3. Integrate at the level of the measurement principle, not the finished product; fusing finished products loses the physics that lets uncertainties be compared. Access to the calibrated instrument measurement is the precondition, and the shared forward models, joint-retrieval frameworks, and expertise behind it would be planned and funded as mission infrastructure, alongside the hardware rather than after it — which is what makes integration a capability rather than a commitment. The same declaration is also what lets a commercial or international dataset be assessed on its measurement chain rather than case by case.
4. Set observing priorities with operational users, and let the user own the threshold of good enough to act on: deliver the number and an honest error bar, and the room to weigh it.
5. Derive the next measurement rather than advocate for it. What the joint system cannot see, stated as a quantified requirement in physical units, is a case every measurement community can support.

Acknowledgments

A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Disclosure of AI use. The lead author used Claude (Anthropic) to refine the English of this white paper so that the points are conveyed clearly. The ideas, arguments, and recommendations are solely the authors' own.

References

Bohn, N., et al. (2021). Optimal estimation of snow and ice surface parameters from imaging spectroscopy measurements. *Remote Sensing of Environment*, 264, 112613. DOI: 10.1016/j.rse.2021.112613

Bonnell, R., et al. (2024). L-band InSAR snow water equivalent retrieval uncertainty increases with forest cover fraction. *Geophysical Research Letters*, 51, e2024GL111708. DOI: 10.1029/2024GL111708

Braverman, A., Hobbs, J., Teixeira, J., and Gunson, M. (2021). Post hoc uncertainty quantification for remote sensing observing systems. *SIAM/ASA Journal on Uncertainty Quantification*, 9(3), 1064–1093. DOI: 10.1137/19M1304283

Dorigo, W., et al. (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions. *Remote Sensing of Environment*, 203, 185–215. DOI: 10.1016/j.rse.2017.07.001

Entekhabi, D., et al. (2010). The Soil Moisture Active Passive (SMAP) mission. *Proceedings of the IEEE*, 98(5), 704–716. DOI: 10.1109/JPROC.2010.2043918

Green, R. O., et al. (2020). The Earth Surface Mineral Dust Source Investigation: An Earth science imaging spectroscopy mission. *IEEE Aerospace Conference*, 1–15. DOI: 10.1109/AERO47225.2020.9172731

Gruber, A., et al. (2020). Validation practices for satellite soil moisture retrievals: What are (the) errors? *Remote Sensing of Environment*, 244, 111806. DOI: 10.1016/j.rse.2020.111806

Guneriussen, T., Høgda, K. A., Johnsen, H., and Lauknes, I. (2001). InSAR for estimation of changes in snow water equivalent of dry snow. *IEEE Transactions on Geoscience and Remote Sensing*, 39(10), 2101–2108. DOI: 10.1109/36.957273

Kellogg, K., et al. (2020). NASA-ISRO Synthetic Aperture Radar (NISAR) mission. *IEEE Aerospace Conference*, 1–21. DOI: 10.1109/AERO47225.2020.9172638

National Academies of Sciences, Engineering, and Medicine [NASEM] (2018). *Thriving on Our Changing Planet: A Decadal Strategy for Earth Observation from Space.* The National Academies Press. DOI: 10.17226/24938

National Academies of Sciences, Engineering, and Medicine [NASEM] (2026). *Decadal Survey for Earth Science and Applications from Space 2028–2037: Request for Information.* https://www.nationalacademies.org/our-work/decadal-survey-for-earth-science-and-applications-from-space-2028-2037

O, S., and Orth, R. (2021). Global soil moisture data derived through machine learning trained with in-situ measurements. *Scientific Data*, 8, 170. DOI: 10.1038/s41597-021-00964-1

O'Neill, P. E., et al. (2021). *SMAP Algorithm Theoretical Basis Document: Level 2 & 3 Soil Moisture (Passive) Data Products*, Revision G. Jet Propulsion Laboratory, JPL D-66480.

Painter, T. H., et al. (2016). The Airborne Snow Observatory: Fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. *Remote Sensing of Environment*, 184, 139–152. DOI: 10.1016/j.rse.2016.06.018

Rasmussen, R., et al. (2012). How well are we measuring snow? The NOAA/FAA/NCAR winter precipitation test bed. *Bulletin of the American Meteorological Society*, 93(6), 811–829. DOI: 10.1175/BAMS-D-11-00052.1

Rodgers, C. D. (2000). *Inverse Methods for Atmospheric Sounding: Theory and Practice.* World Scientific. DOI: 10.1142/3171

Serreze, M. C., et al. (1999). Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resources Research*, 35(7), 2145–2160. DOI: 10.1029/1999WR900090

Singh, S., et al. (2024). Bayesian physical–statistical retrieval of snow water equivalent and snow depth from X- and Ku-band synthetic aperture radar — demonstration using airborne SnowSAR in SnowEx'17. *The Cryosphere*, 18, 747–773. DOI: 10.5194/tc-18-747-2024