The Astrophysical Journal, 970:168 (29pp), 2024 August 1 https://doi.org/10.3847/1538-4357/ad41e3

SuperSynthIA: Physics-ready Full-disk Vector Magnetograms from HMI, Hinode, and Machine Learning

Ruoyu Wang[1], David F. Fouhey[1,2,3], Richard E. L. Higgins[3], Spiro K. Antiochos[4], Graham Barnes[5], J. Todd Hoeksema[6], K. D. Leka[5], Yang Liu[6], Peter W. Schuck[7], and Tamas I. Gombosi[8]
[1] New York University, Courant Institute of Mathematical Sciences, New York, NY, USA; david.fouhey@nyu.edu
[2] New York University, Tandon School of Engineering, New York, NY, USA
[3] University of Michigan, Department of Electrical Engineering and Computer Science, Ann Arbor, MI, USA
[4] University of Michigan, Department of Climate and Space Sciences and Engineering, Ann Arbor, MI, USA
[5] NorthWest Research Associates, Boulder, CO, USA
[6] Stanford University, Stanford, CA, USA
[7] Heliophysics Science Division, NASA Goddard Space Flight Center, 8800 Greenbelt Rd., Greenbelt, MD 20771, USA
[8] University of Michigan, Department of Climate and Space, Center for Space Environment Modelling, Ann Arbor, MI, USA
Received 2023 December 6; revised 2024 February 26; accepted 2024 March 22; published 2024 July 29

Abstract

Vector magnetograms of the Sun's photosphere are cornerstones for much of solar physics research. These data are often produced by data-analysis pipelines combining per-pixel Stokes polarization vector inversion with a disambiguation that resolves an intrinsic 180° ambiguity. We introduce a learning-based method, SuperSynthIA, that produces full-disk vector magnetograms from Stokes vector observations. As input, SuperSynthIA uses Stokes polarization images from Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI). As output, SuperSynthIA simultaneously emulates the inversion and disambiguation outputs from the Hinode/Solar Optical Telescope-Spectro-Polarimeter (SOT-SP) pipeline. Our method extends our previous approach SynthIA with heliographic outputs as well as using an improved data set and inference method. SuperSynthIA provides a new tool for improved magnetic fields from full-disk SDO/HMI observations using information derived from the enhanced capabilities of Hinode/SOT-SP. Compared to our previous SynthIA, SuperSynthIA provides physics-ready vector magnetograms and mitigates unphysical angle preferences and banding artifacts in SynthIA. SuperSynthIA data are substantially more temporally consistent than those from the SDO/HMI pipeline, most notably seen in, e.g., evolving active regions. SuperSynthIA substantially reduces noise in low-signal areas, resulting in less center-to-limb bias outside of strong-signal areas. We show that outputs from SuperSynthIA track the SDO/HMI-recorded evolution of the magnetic field. We discuss the limitations of SuperSynthIA that the user must understand, and we demonstrate a broad set of evaluations to test SuperSynthIA and discuss remaining known artifacts. Our tests provide both methodology and evidence that SuperSynthIA outputs are ready for use by the community, and that learning-based approaches are suitable for physics-ready magnetograms.

Unified Astronomy Thesaurus concepts: Solar magnetic fields (1503); Convolutional neural networks (1938); Computational methods (1965)

Supporting material: animation

1. Introduction

High-cadence, accurate, and physically consistent estimates of the Sun's photospheric magnetic field serve as the cornerstone for much of solar physics, from estimating the static configuration of the coronal magnetic field (Wiegelmann & Sakurai 2021) to modeling the physical evolution of the Sun (Cheung & DeRosa 2012; Lionello et al. 2014; Gombosi et al. 2018; Hayashi et al. 2021), to understanding and forecasting space weather (Bobra et al. 2014; Barnes et al. 2016; Leka et al. 2018). Accordingly, a host of instruments with varying capabilities for inferring the solar magnetic field has been deployed, including the Helioseismic and Magnetic Imager (HMI; Schou et al. 2012) on the Solar Dynamics Observatory (SDO; Pesnell et al. 2012), and the Solar Optical Telescope-Spectro-Polarimeter (SOT-SP; Tsuneta et al. 2008) on the Hinode mission (Kosugi et al. 2007). Accompanying these physical platforms has been extensive effort to automatically convert the observations of polarized light into physically accurate estimates of the photospheric vector magnetic field. These pipelines incorporate simplifying assumptions, calibration, data preparation, and various implementation details. Substantial effort over the past two decades has gone into both the instrumentation and the software pipelines to produce magnetograms with two characteristics: first, recovering the true *vector* field; and second, having relatively few systematic artifacts.

Recovering the true vector field, and in particular all three components of the vector field, is important because many applications, such as estimating the free energy and magnetic helicity flux into an active region (e.g., Gary et al. 1987; Pevtsov et al. 1994; Liu & Schuck 2012; Schuck & Antiochos 2019), require an unambiguous, physically meaningful vector. In addition to the challenge of inferring the line of sight and transverse components of the field and its host atmosphere from the polarization spectra (Centeno et al. 2014; del Toro Iniesta & Ruiz Cobo 2016), the field component transverse to the line of sight has an intrinsic 180° ambiguity (Harvey 1969) that can

require consideration of multiple—typically many—pixels to fully resolve (see Metcalf et al. 2006, for a review). This additional step, plus the transform to recover the physical magnetic field components (Gary & Hagyard 1990), is required in order to perform calculation of many quantities of interest as alluded to above (see, e.g., Bobra et al. 2014; Hoeksema et al. 2020, and references therein).

Removing artifacts is also critical because their presence can fundamentally limit or preclude downstream applications. Consider, for instance, the use of magnetograms as boundary conditions for magnetohydrodynamic (MHD) models. If temporal inconsistency creates the appearance of frequent high-speed reconfigurations of the magnetic field, models must contort to explain away these phantom changes. All of the above will limit the ability to accurately infer the buildup of coronal free energy and helicity, which are critically important quantities for any space weather prediction capability.

In this paper, we introduce and analyze SuperSynthIA, a machine learning (ML)-based approach for estimating the full-disk, disambiguated photospheric vector magnetic field from the Stokes vector. This work builds on our prior effort SynthIA (Higgins et al. 2022): a method is trained with observations of one instrument as an input and the pipeline results of another instrument as an output, enabling the method to inherit advantages of both instruments. In our case, we learn a mapping from SDO/HMI Stokes polarization vector observations to vector magnetic field products from Hinode/SOT-SP, which takes advantage of the instruments' complementary strengths. As an input, SDO/HMI Stokes vectors are captured as a coherent full-disk image with good sampling (0.″5) and regular rapid cadence (e.g., 12 minutes for standard data products), enabling a monumental near-complete full-disk archive spanning more than a decade. Hinode/SOT-SP Stokes data, on the other hand, captures a narrow field of view over the course of 30 minutes to an hour (with polar scans even longer) due to its scanning slit spectrograph mechanism. Moreover, due to the narrow field of view and mechanism, targets of interest may not be captured continuously, at high cadence, at a time of interest, or at all (unlike SDO/HMI's continuous, full visible disk, high-cadence observations). Hinode/SOT-SP's longer acquisition time, however, enables the capture of two spectral lines with different Landé factors as opposed to SDO/HMI's single line, and 112 spectral samples as compared to SDO/HMI's 6. Accordingly, the resulting HinodeP output data are more likely to better capture the state of the photosphere. Once learned, SuperSynthIA does not need HinodeP data at inference time, and can produce Hinode/SOT-SP-like data over the full-disk from SDO/HMI Stokes images—bringing the possibility of Hinode/SOT-SP-like vector data to the full-disk at a rapid cadence over SDO/HMI's entire mission. As an additional benefit, due to its use of machine learning, SuperSynthIA is extremely fast and, as we show later, temporally consistent.

The hypothesis underlying this line of work is that the information needed for Hinode/SOT-SP-like data are present in the SDO/HMI Stokes data but potentially hidden over multiple pixels, and that this information can be extracted using supervised machine learning. We see the ML approach as complementary to work aimed at analytically identifying this information, such as by Griñón-Marín et al. (2021) who developed an alternate inversion pipeline for SDO/HMI that can provide a magnetic fill factor. The ML-based approach has the advantage that only pairs of inputs and outputs are needed, and therefore, one does not need an explicit physical model.

The data products used in this work are the result of complex pipelines, and so, before making concrete claims, some precision in language is needed. The vector field products used are the results of the following: instruments and their calibration (for obtaining Stokes spectropolarimetric data); Milne–Eddington (ME; Unno 1956; Rachkovsky 1962) inversion algorithms (that recover the image-plane field components, caveat the 180° ambiguous transverse component $B_\perp$ from the polarization spectra); and disambiguation algorithms (that choose between the two angles for $B_\perp$). Telescope pointing information is also used, and is particularly important for producing the vector components in heliographic coordinates. In this paper, we use "HMIP"—the HMI Pipeline—to refer to the ME inversion data products produced by the pipeline variant of VFISV (Borrero et al. 2011; Centeno et al. 2014) and the disambiguated data products produced via a variant of the minimum-energy (ME0) disambiguation procedure ("ME0"; Metcalf 1994; Hoeksema et al. 2014). Both originate from the SDO/HMI Stokes filtergrams. Similarly, we use "HinodeP" to refer to the pipeline ME inversion products produced by MERLIN (Lites et al. 2007) from Hinode/SOT-SP's Stokes polarization spectra as well as the same ME0 disambiguation (albeit with slightly different implementation due to data and timing constraints). To avoid more terminology, we also use HinodeP to refer to the HinodeP data remapped to SDO/HMI's grid, as is done throughout this paper.

Throughout, we focus our efforts on estimates of the strength of the field and its angles due to their ubiquitous use. Unlike our previous works, our outputs include the four properties available after disambiguation: the heliographic components of the magnetic flux density vectors αB_R, αB_ϕ, αB_θ as well as the disambiguated azimuth ψ_{360}. These are illustrated in Figure 1, which highlights the key outputs provided by SuperSynthIA. We additionally provide five properties obtained from an ME inversion: intrinsic field strength B, fill fraction α, their product αB, as well as plane-of-sky inclination γ, and azimuth ψ angles.

We show that SuperSynthIA is able to provide a close estimate of HinodeP's output as compared to SynthIA as well as provide vector magnetograms that have improvements on several artifacts. Compared to our past method, the inference improvements of SuperSynthIA enable it to avoid preferential angles in low-signal regions and banding artifacts. Temporally, SuperSynthIA shows an order of magnitude reduction in temporal "flickering" of the vector magnetic field, e.g., as evaluated during periods of flux emergence. In low-signal regions, SuperSynthIA substantially suppresses noise; as a result, SuperSynthIA substantially reduces a center-to-limb variation in "quiet Sun" magnetic field that usually needs to be masked out with HMIP while retaining HMIP's good center-to-limb consistency in high-signal regions such as active regions. Over long periods of time, SuperSynthIA matches HMIP's results quite well in long-term full-disk field trends as well as the polar radial field. Of course, no data are perfect, and we discuss several limitations of SuperSynthIA that end users must be aware of in order to arrive to proper scientific interpretations.

As an added benefit, SuperSynthIA produces results quickly in its current form, but still with substantial room for potential improvement. In turn, this speed opens doors to new applications (e.g., using the network as a real-time system). In particular, each magnetic field property can be obtained with

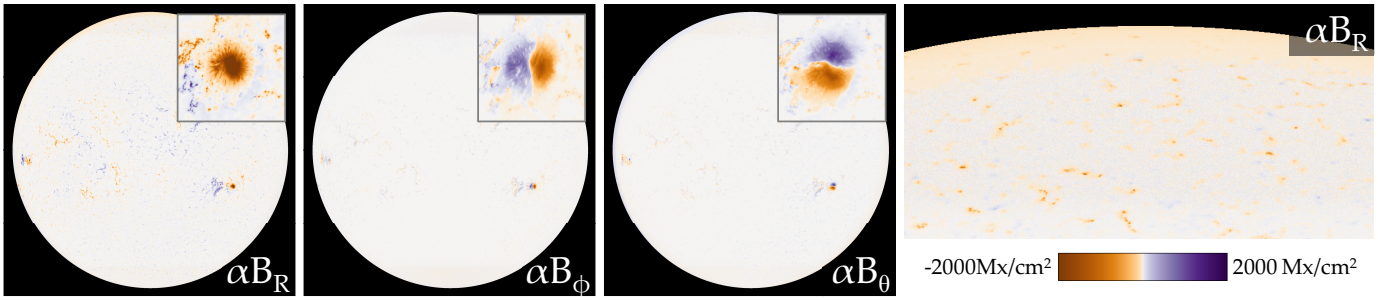

Figure 1. Example outputs from SuperSynthIA include physically relevant data products, such as the heliographic components (left to right) αB_R, αB_ϕ, αB_θ in addition to the standard Milne–Eddington inversion products. On far right, we show a polar region. Outputs from SuperSynthIA closely resemble the output of the Hinode/SOT-SP vector magnetogram pipeline, including cleaner estimates of the field in low-signal regions and reduced presence of spatiotemporal artifacts affecting downstream applications. At the same time, SuperSynthIA produces magnetograms from the full-disk rapid cadence observations of SDO/HMI. Each output of SuperSynthIA can be produced by the forward pass of a deep network, leading to substantial acceleration compared to a traditional Milne–Eddington inversion plus energy-minimization disambiguation. Example date, 2016 February 1, 12:36 TAI; colorbar, -2000 2000 Mx cm^{-2} for αB_R, αB_ϕ, αB_θ. The colormaps for magnetic field/flux quantities in this figure (and all other figures) are *nonlinear*. We take the signed square root ($x \mapsto \mathrm{sign}(x)\sqrt{|x|}$) of each value and the bounds before plotting, which helps better simultaneously show strong and weaker fields.

a forward pass of a deep network, taking less than 30 s on a standard GPU with an unoptimized method. This runtime has *ample* room for acceleration, including standard network compression tricks and training multiple targets simultaneously. However, we see these efforts as orthogonal and a secondary goal to be investigated after first obtaining high accuracy and a good understanding of the strengths and weaknesses of the learning-based approach.

2. Inversion and Disambiguation Data

A full introduction to Stokes vector inversion and disambiguation is beyond the scope of the paper, but we provide a brief description of the standard products of HMIP to make the paper more self-contained and to point out particular properties that will be drawn upon later. The reader is encouraged to read Hoeksema et al. (2014) for a full description of the SDO/HMI data products used and del Toro Iniesta & Ruiz Cobo (2016) for a more detailed description of the general field. Both HMIP and SuperSynthIA use Stokes vectors measured from the `hmi.S_720s` data series that consists of the four Stokes components [I, Q, U, V], sampled with six filtergram passbands.

The Stokes vector inversion process aims to find an atmosphere that is consistent with the observed Stokes vectors; in both the cases of HMIP, using the inversion code VFISV (Borrero et al. 2011), and HinodeP, using the inversion code MERLIN Community Spectropolarimetric Analysis Center (CSAC 2006; Ramos & Baso 2019), pipelines make the simplifying assumption of an ME model, in which the magnetic field and thermodynamic properties are constant across optical depth, and the source function is linear with optical depth. ME models have been shown to generally produce good averages of the magnetic vector across optical depth (Westendorp Plaza et al. 1998; Borrero et al. 2014). Intrinsic to the Zeeman effect underlying the model is a 180° ambiguity in the recovered vector, which a subsequent disambiguation step aims to resolve (see Metcalf et al. 2006).

2.1. Milne–Eddington Stokes Inversion

Our first five targets are standard ME Stokes inversion products:

1. The field strength (B), measured in gauss, is the intrinsic field strength in the pixel. HinodeP reports this as `Field_Strength`, and separates field strength B from the fraction of the pixel that is occupied by magnetized plasma. HMIP does not separate the two, but instead reports the ME results assuming field-filled pixels.
2. The fill factor (α) is the unitless quantity representing the fraction of the pixel that is filled with a magnetic field. HinodeP reports this as `Stray_Light_Fill_Factor`. HMIP does not independently fit for this parameter, and instead assumes that $\alpha = 1$, which is likely mostly correct in deep sunspot umbrae, but is not a good assumption elsewhere, even at the higher Hinode/SOT-SP spatial resolution. It should be noted that Griñón-Marín et al. (2021) have developed a model capable of inferring fill factor for SDO/HMI data, at the cost of an increased runtime that ranges from 1.6× HMIP if only applied to pixels with total polarization $\geqslant$0.25%, to 5× for all pixels (Liu et al. 2022).
3. The product of field strength and fill factor (αB) is thus the common quantity between the two instruments and a focus of our analysis here. To distinguish this product from the intrinsic field strength (with units gauss), αB is often reported with units Mx cm^{-2} to include the implication of the geometric factor α being included, whether assigned (as with HMIP as `field`), or variable (as with HinodeP) but reported as their product. HMIP reports this as `field`; we calculate αB for HinodeP as the product of α and B before interpolation.
4. The inclination angle (γ), measured in degrees and ranging from 0° (toward the observer) to 180° (away from the observer), is used to determine the strength of both the line-of-sight magnetic field $\alpha B_\parallel = \cos(\gamma)\alpha B$ and the transverse magnetic field $\alpha B_\perp = \sin(\gamma)\alpha B$. Noise in the positive-definite $B_\perp$ can lead to preferred inclinations close to 90° in regions with a low-polarization signal (LaBonte 2004; Pevtsov et al. 2021).
5. The azimuth (ψ) is measured in degrees and ranges from 0° to 180°. The ambiguity is a property of the Zeeman effect; two atmospheres with identical parameters except for a 180° change in azimuth orientation (e.g., $\psi = 45°$ and $\psi = 225°$) produce indistinguishable Stokes profiles.

2.2. Ambiguity-resolved and Heliographic Vector Quantities

The inherent 180° ambiguity in the inferred plane-of-sky azimuth specifies two vectors. Accordingly, a critical next step is disambiguation to assign which vector is more likely. Disambiguation can be seen as identifying a *disambiguation bit* for each pixel that determines whether the azimuth lies within [0°, 180°) or [180°, 360°). In the sophisticated algorithms, this bit is determined by way of information from multiple pixels (see Metcalf et al. 2006, for discussion) through modeling or invoking the expected behavior that relies on spatial derivatives. For the Minimum Energy algorithm, a global optimization minimizes a function that incorporates both vertical current density and an estimate of the field divergence (Metcalf 1994; Leka et al. 2009; Hoeksema et al. 2014). Once the disambiguation bit is identified, the heliographic magnetic field vector is determined following Gary & Hagyard (1990).

There are many ways to identify the disambiguation bit. Both HinodeP and HMIP use the Minimum Energy (Metcalf 1994) algorithm. HinodeP uses ME0 at every pixel due to the considerably smaller data volume. For speed, HMIP uses ME0 only for pixels with sufficiently strong transverse field $B_\perp$, and provides an assigned disambiguation bit to weak-field pixels with three options: acute-angle to a potential field, acute-angle to the radial direction, or randomized outcomes. Unless otherwise specified, we use the random option since it shows no systematic biases, unlike the others.

Once the bit has been identified, it defines a few extra quantities that can then be directly calculated:

1. Azimuth360 (ψ_{360}), measured in degrees and ranging from 0° to 360°, is the disambiguated azimuth. Given a disambiguation bit $b \in \{0, 1\}$, ψ_{360} is calculated as $\psi + 180b$. The disambiguated azimuth ought to not show preferential directions in regions with a weak transverse field, although the strictest test is a lack of bias in noise-dominated areas in the resulting heliographic components.
2. αB_R is the radial component of the magnetic flux density in heliographic coordinates and measured in $\mathrm{Mx\,cm^{-2}}$. Positive is oriented against gravity; negative is oriented toward the Sun's center. The radial component is derived from αB, the magnetic field vector angles γ, ψ, disambiguation bit, and location on the Sun.
3. αB_ϕ is the poloidal component that aligns along longitudes and also measured in $\mathrm{Mx\,cm^{-2}}$. Positive is oriented toward west, and negative toward east.
4. αB_θ is the toroidal component that aligns along latitudes and also measured in $\mathrm{Mx\,cm^{-2}}$. Positive is oriented toward north and negative toward south.

We obtain the HMIP data products from JSOC in the form of the `hmi.ME_720s_fd10` (ME inversion) and `hmi.B_720s` (disambiguated) data series. We obtain the HinodeP data from the Community Spectropolarimetry Analysis Center (Ramos & Baso 2019) and use the Level 2 (ME inversion) and Level 2.1 (disambiguated) products.

2.3. Artifacts

The data products involved have known artifacts, and different instruments and pipelines have different known artifacts. Due to the process for generating vector magnetograms, an artifact impacting field-strength-derived quantities such as αB will also impact the heliographic components such as αB_R. Several of the artifacts, especially ones in HMIP, are known to the community and are well understood.

In *low-signal* pixels, HMIP shows a spatial viewing angle bias consisting of an increase in αB from the center to the limb. This bias is *not* present in well-measured pixels, such as in active regions, at least for data more than 15°–20° degrees from the limb. We refer readers to the discussion of Hoeksema et al. (2014). This bias needs to be accounted for in data analysis using the `conf_disambig` mask that identifies pixels having sufficient signal to be appropriate for further data analysis. The bias is best seen via an increase in the nominal quiet Sun field strength (as well as noise) toward the limb, including the poles.

There are temporal oscillations in HMIP data including in the field and other quantities. The average αB over the disk in HMIP displays an oscillation that is most clearly visible at a 24 hr frequency, but also appears in its harmonics (Hoeksema et al. 2014). The artifact is believed to arise from a mismatch between model and in-flight optical elements that, when coupled with the large ($\pm 3\,\mathrm{km\,s^{-1}}$) relative velocities of the geosynchronous orbit, lead to an imperfect calibration of the polarization spectra as a function of orbit, and results in substantial changes in the inferred magnetic fields.

Finally, there is a so-called "hemispheric bias" (but more precisely, an $\alpha B_\perp$ bias) in HMIP, as discussed in Pevtsov et al. (2021), Liu et al. (2022), Griñón-Marín et al. (2021), Leka et al. (2022) that is substantially exacerbated by the assumption of a fill factor of unity in HMIP. This bias causes the inferred heliographic direction of the magnetic field to change unphysically as a target transits across the disk, especially targets that are intrinsically spatially unresolved (e.g., plage).

While HinodeP is not believed to have the 24 hr oscillation or hemispheric bias, it does have other artifacts. The heliographic components of HinodeP show artifacts at a few dozen SDO/HMI pixels at the limb. This is because heliographic components depend on not just inversions and disambiguation, but also the location on the Sun (Gary & Hagyard 1990). This pointing information is known to be inaccurate with a directional bias (Fouhey et al. 2023), leading to inferred near- and off-limb artifacts. This artifact must be addressed by reprocessing all of the HinodeP data with proper pointing information and is beyond the scope of this article.

Finally, there are machine learning-created artifacts, including preferential directions in weak fields. When $\alpha B_\perp$ is weak and dominated by noise, the azimuth should have no preferred direction. Optimization-based Stokes inversion approaches like the methods used to produce HMIP or HinodeP do not show preferences, but Higgins et al. (2022) show directional preference in weak-field regions that can be seen as large regions with the same value. We discuss the origin of this phenomenon and mitigations subsequently.

3. Data Alignment

Learning-based systems depend intrinsically on the quality of the inputs and outputs. Our earlier work (Higgins et al. 2022) presented a proof of concept version that we expand here. Alignment is difficult for two reasons. First, the reported pointing information of Hinode/SOT-SP is known to be inaccurate (Hurlburt et al. 2009, Appendix C), which our previous work (Fouhey et al. 2023) suggests derives from issues with Hinode's Attitude and Orbit Control System. The lack of accurate pointing information means that one cannot

synthesize a warping field between the two observations (as in other cross-instrument analyses). Second, even if the pointing information *were* accurate, the 30 minute to hour-long acquisition process for a Hinode/SOT-SP scan means that substantial evolution happens during the observation. Thus, no one *rigid* registration can accurately align SDO/HMI and Hinode/SOT-SP.

Our goal is to fit an alignment transformation that describes a mapping from SDO/HMI pixel coordinates $[x, y]^\top$ to Hinode/SOT-SP pixel coordinates $[x', y']^\top$. Since Hinode/SOT-SP is a scanning slit spectrograph, the columns of its Level 2 magnetograms indicate temporal *indices* not spatial *positions*. We accordingly use the `Mechanical_Slit_Position` field in the Hinode/SOT-SP data to expand the magnetogram so that every *position* has image data. We linearly interpolate data at those spatial positions that were not observed during the scan. Since spatial jumps are fairly rare, this usually interpolates only a handful of spatial positions.

This alignment transformation consists of two components, a rigid transformation between Hinode/SOT-SP and SDO/HMI and a non-rigid flow-field correction. The rigid transformation consists of a rotation, anisotropic scaling, and translation. Following Fouhey et al. (2023), we describe this as mapping from Hinode/SOT-SP to SDO/HMI pixels (although note that the notation differs from Fouhey et al. 2023 in whether Hinode/SOT-SP or SDO/HMI pixels are x, y or x', y'). The rigid transformation $f_R([x', y']^\top)$ takes the form

$$\boldsymbol{R}(\theta, c_x, c_y)\begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix}\begin{bmatrix} x' \\ y' \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{1}$$

where $\boldsymbol{R}(\theta, c_x, c_y)$ is a rotation by θ around a center point c_x, c_y. This function f_R is invertible, and f_R^{-1} is a mapping from SDO/HMI to Hinode/SOT-SP coordinates. The relative scalings s_x and s_y are held fixed using the relative scalings determined by Fouhey et al. (2023), which are nearly the nominal scalings given by the Hinode/SOT-SP Level 1 pixel scales. We solve for t_x, t_y, and θ using two distinct strategies, one for equatorial scans and the other for polar scans.

The nonrigid flow $f_N([x, y]^\top)$ shifts the SDO/HMI coordinates to account for temporal evolution. The flow takes the form $[x + \boldsymbol{U}[x, y], y + \boldsymbol{U}[x, y]]^\top$ and is parameterized by flow maps $\boldsymbol{U}$ and $\boldsymbol{V}$ that contain, for each pixel (x, y), a potentially subpixel flow (i.e., $\boldsymbol{U}[x, y] \in \mathbb{R}$). This flow map is almost certainly not invertible, but the lack of an inverse is not an issue since we will warp the Hinode/SOT-SP data onto the SDO/HMI grid only.

The final transformation is a mapping from SDO/HMI to Hinode/SOT-SP coordinates. This is a composition of the small nonrigid shift followed by a rigid transformation, $f([x, y]) = f_R^{-1}(f_N([x, y]))$. We fit the rigid transformation differently depending on whether the scan is equatorial or polar, described below. Two different methods are needed due to the difference in available cues for alignment. The equatorial scans often contain texture that is easily aligned with standard computer vision techniques, e.g., Lowe (2004) but no limb. Polar scans show a limb, but limited texture, and so need a different method.

3.1. Alignment of Equatorial Scans

We automatically identify scans as either equatorial or polar. Since identification is automatic, some errors may occur. Mistakes tend to be the inclusion of high-latitude limb scans in the polar scans; to the best of our knowledge, there are no polar scans in the equatorial scans. To align equatorial scans, we follow Fouhey et al. (2023) and register via SIFT (Lowe 2004) and ORB (Rublee et al. 2011). For full details, we refer readers to our previous work. However, the key details are that a parametric model is robustly fit to automatically find correspondences between HinodeP and HMIP data using a well-understood computer vision approach (Hartley & Zisserman 2004; Snavely et al. 2006) modified for use on a scanning slit spectrograph.

3.2. Alignment of Polar Scans

Aligning poles is challenging due the magnetic structure of the poles and the length of time required for a Hinode/SOT-SP polar scan. The polar magnetic structure is predominantly low-signal quiet areas punctuated by occasional strong small features, and overall a severe lack of large, strong texture features. We solve for t_x, t_y, and θ in two stages.

First, we align the disks by solving for t_x and t_y. Since the reported pointing information for Hinode/SOT-SP is inaccurate and thus cannot produce a disk mask, we create one from the inferred continuum intensity (`Continuum_Intensity`< 2500), total χ^2 field (`ChiSq_Total` $>10^{-5}$), and the inferred gradient of the source function (`Source_Function_Gradient` < 5000). We take a majority vote and apply basic morphological operators (`binary_fill_holes` from Scipy followed by a 5×5 binary erosion). We then cross-correlate this disk with a disk mask from SDO/HMI, yielding a translation. While the line formation heights for the observations of SDO/HMI and Hinode/SOT-SP are different, the analysis presented in Fouhey et al. (2023) suggests that the difference is less than an SDO/HMI pixel.

We then find the rotation θ with the t_x, t_y fixed by maximizing alignment between αB maps. Due to the noise characteristics of SDO/HMI, the two instruments disagree about both the weak polarization-signal regions and the strength of the strong-field regions. However, they agree on the *locations* of the strong field. We thus apply a truncated rank transformation that functions like a softened strong-field mask. We replace each pixel with its percentile (i.e., $[370, 60, 442, 212] \rightarrow [\frac{2}{3}, 0, 1, \frac{1}{3}]$) to use the *relative ordering* of the strengths. We then truncate the percentile to the top 5% of the data or clip $((p - 0.95)/0.05, 0, 1)$ where p is the percentile. The truncation ignores the weak-field regions, and is done since the pipelines differ in their interpretation of these regions.

Given a rotation, we compute disagreement between the rank-transformed data via weighted sum. We perform a two-level coarse-to-fine brute-force search to minimize disagreement: we find an initial solution by searching within $\pm 2^\circ$ of 0 in steps of $0^\circ.04$; we then search $\pm -0^\circ.2$ around the initial solution in steps of $0^\circ.004$. One resulting alignment is shown in Figure 2.

3.3. Nonrigid Registration with Optical Flow

In both cases, there is residual misalignment between the images left, so we perform optical flow using the approach of Liu (2009), using the wrapper from Pathak et al. (2017). Initially, we explored learning-based approaches, but classic methods are appealing *precisely* because they are not learned but instead minimize a defined objective function. This

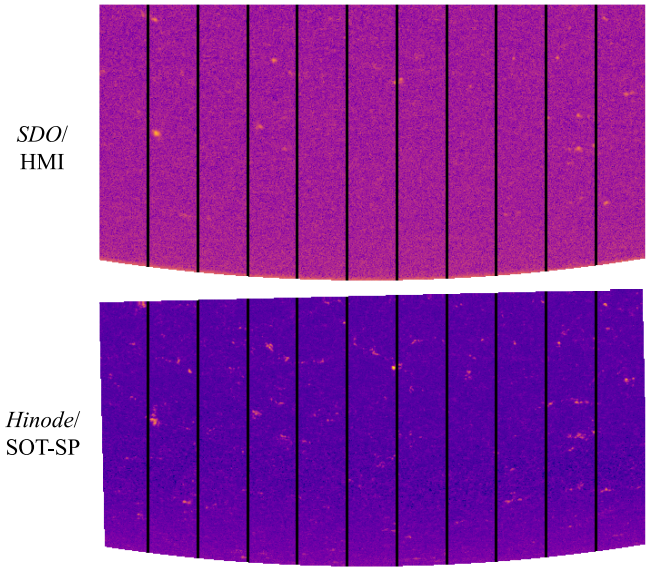

Figure 2. An aligned polar scan. The inferred αB for Hinode/SOT-SP is substantially different from HMIP, complicating alignment. Alignment is based on a truncated rank-transform that captures which pixels are strongest (rather than their strength). As a result, the poles are well aligned, although there is often a small amount of remaining misalignment due to the (often) multihour acquisition time. Colormap: 0 1000 Mx cm^{-2}. Example from: 2012 January 13, 13:02:52 TAI.

objective function often generalizes to domains without good learned correspondence models.

Given an initial rigid transformation, we compute a nonrigid flow correction on the αB maps. Let $\boldsymbol{I}_1$ denote the αB from SDO/HMI, specifically the `field` map from the `hmi.ME_720s_fd10` series. Then, let $\boldsymbol{I}_2$ denote the Hinode/SOT-SP αB data warped using the initial rigid transformation. The αB is the per-pixel product of the Hinode/SOT-SP Level 2 `Stray_Light_Fill_Factor` and `Field_Strength`.

The flow approach of Liu (2009) has two key terms that it uses to find an optimal optical flow field $\boldsymbol{U}$, $\boldsymbol{V}$. The first is a data agreement term

$$\sum_{x,y} |\boldsymbol{I}_1(x + \boldsymbol{U}[x, y], y + \boldsymbol{V}[x, y]) - \boldsymbol{I}_2[x, y]| \tag{2}$$

that measures how well the images align after warping in the ℓ_1 sense. The second is a smoothness term on the flow vectors, taking the form of the norm of the gradient

$$\sum_{x,y} (||\nabla \boldsymbol{U}[x, y]||^2 + ||\nabla \boldsymbol{V}[x, y]||^2)^{0.5}. \tag{3}$$

There are additional terms that aim to enforce symmetry, and the objectives are minimized in a coarse-to-fine fashion. We refer the reader to Liu (2009). The data agreement and smoothness terms are given a relative weighting; this weighting is 5× the value for the poles because the data intrinsically are substantially different. Not enforcing smoothness leads to noisy flow that attempts to fix the substantial differences in appearance by moving pixels around implausibly.

3.4. Warping

Given the mapping $f([x, y])$ from SDO/HMI pixels to Hinode/SOT-SP pixels, we produce aligned Hinode/SOT-SP-like image data on the SDO/HMI grid. Discrete quantities such as flags are handled via nearest-neighbor interpolation. Continuous quantities like field strength are warped onto the grid using bilinear interpolation. We chose bilinear interpolation since it avoids ringing artifacts that a cubic spline or other higher-order interpolation would cause. Ringing would create physically impossible values such as negative field strength. Finally, circular quantities such as azimuthal angles are handled by interpolating in 2D and reprojecting to avoid preferential directions in interpolations. The angle θ is mapped to $[\sin(2\theta), \cos(2\theta)]$ for 180° ambiguous quantities and to $[\sin(\theta), \cos(\theta)]$ for other circular quantities. Following interpolation, the results are mapped back to angles with the two-argument arctangent.

We include a 100 pixel (SDO/HMI pixels) buffer around each complete HinodeP scan that is aligned to the SDO/HMI grid. The result is illustrated in Figure 3 where the buffer region is shown in gray. In this ≈50″ buffer region, both SDO/HMI Stokes data and HMIP vector magnetogram data are available, but HinodeP data are not. We include this buffer data with the aim of improving machine learning approaches. The decision of a deep learning model depends on a larger region than a pixel (known as a receptive field). SuperSynthIA can use this additional context to produce disambiguated results. By providing padding in which SDO/HMI Stokes data are available, all pixels on which training is done receive meaningful context.

3.5. Data Characteristics

We use the above procedures to align every Hinode/SOT-SP scan between 2011 January 1 and 2021 December 31 with all contemporaneous SDO/HMI scans. This data volume is substantially larger than the 2 yr one used in Higgins et al. (2022), comprising 13.4K Hinode/SOT-SP scans.

We plot information about the location and observer velocity of SDO/HMI at the time of all target Hinode/SOT-SP scans in Figure 4. The distributions of locations look similar to the data in Higgins et al. (2022) apart from the newly added polar scans. There is substantial sampling over observer velocity as exhibited by the relative smoothness of the plot.

Examples from the data preparation pipeline appear throughout the paper. The warped data looks like the original Hinode/SOT-SP pipeline with a few small differences due to the interpolation that brings Hinode/SOT-SP data to the SDO/HMI grid. First, there is some amount of smoothing due to the rebinning of data. Second, some data are slightly pulled toward the mean (e.g., field strength is decreased), especially for extremely high values. This decrease is due to interpolation averaging over multiple values, which intrinsically decreases the maximum. Finally, Hinode/SOT-SP infers small pixel-sized hot spots in quiet regions that are below SDO/HMI resolution; since the SDO/HMI grid is coarser, these small hotspots are reduced.

4. Method

SuperSynthIA follows the general approach of Higgins et al. (2021, 2022) with a number of modifications that are critical for mitigating unphysical preferences and accurately capturing details. The approach estimates a distribution at each pixel over a sequence of nominal "bin" values. Producing a distribution has several advantages. First, the network produces an estimate of uncertainty that can then be propagated to downstream tasks.

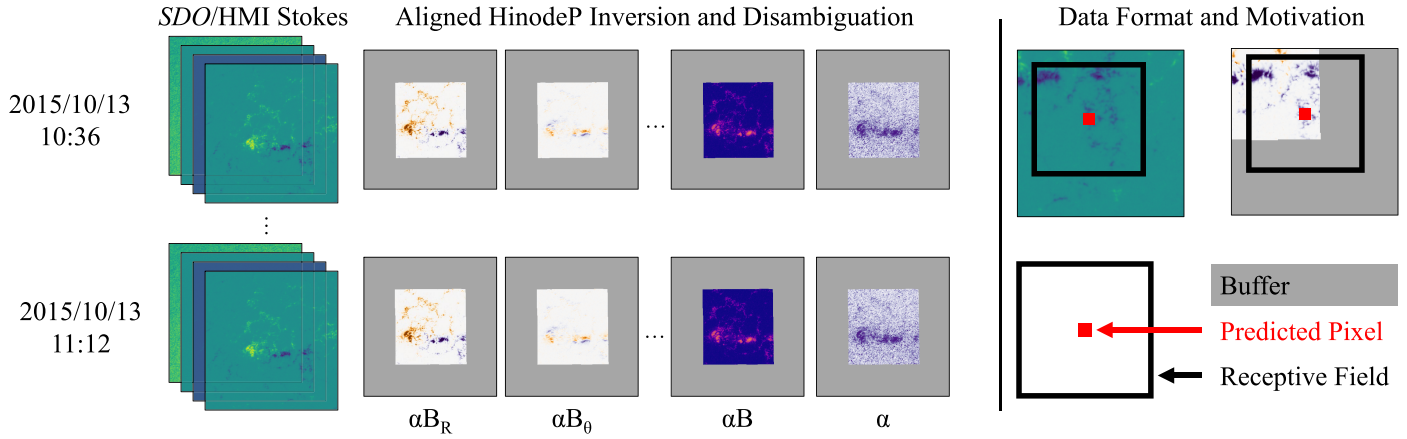

Figure 3. Data format for the data. (Left): for each Hinode/SOT-SP scan, we coalign multiple contemporaneous SDO/HMI scans. These scans are coaligned and matted with a buffer that is 100 HMIP pixels, or ≈50″ (shown in gray). In the buffer, SDO/HMI Stokes observations are available, but HinodeP data are not. (Right): the additional buffer is important because our model's decisions at a pixel (shown in red) depend on a larger region (shown in black), often referred to as the receptive field of the neuron. Providing Stokes observations in the buffer region ensures that all pixels used for training have at least 100 HMIP pixels worth of context (e.g., to help determine disambiguation).

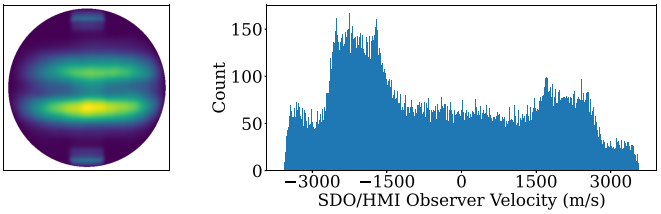

Figure 4. Where and how the scans are taken. (Left): a kernel density plot of the observation longitude/latitude, visualized on an example disk. The asymmetry in terms of density arises from the scan acquisition of Hinode/SOT-SP, in which a region must be selected in advance for imaging. (Right): a histogram of `OBS_VR`, the SDO observer velocity at disk center at the time of capture.

Second, in the setup of SuperSynthIA, where there may be some residual misalignment between the input and target value, we find that it enables better performance compared to regression since the model accounts for outliers by increasing the probability of an outlier bin rather than pulling the regression output toward the outlier.

Our model consists of a parameterized function that maps from an H pixel tall and W pixel wide Stokes vector image input $\boldsymbol{I} \in \mathbb{R}^{H\times W\times N}$ ($N = 24$ for SDO/HMI, comprising the four Stokes vector components at six passbands) to the unnormalized log-likelihoods (logits) for a set of K bin values for each pixel, or $\boldsymbol{Z} \in \mathbb{R}^{H\times W\times K}$. The bin values $\boldsymbol{b} \in \mathbb{R}^{K}$ define the locations over which the network produces a probability distribution. The function itself consists of convolutions interleaved with pointwise nonlinearities, and follows a U-Net-style architecture (Ronneberger et al. 2015) in which the network's spatial resolution contracts, and then reexpands. One critical design decision is that we remove BatchNorm (Ioffe & Szegedy 2015). BatchNorm adaptively normalizes the data per-input channel during training, which is bad for Stokes inversion since the absolute intensity contains crucial information about the field.

Throughout, we follow the architecture from Higgins et al. (2021, 2022) to facilitate fair comparisons. We do not include continuum intensity (SDO/HMI's `hmi.Ic_720s` series). This product is constructed from the I component of the Stokes vector, and so, its information ought to be present already in the Stokes input. This architecture, in unoptimized code, takes under 30 s for a 4096×4096 SDO/HMI observation; one key current limitation is memory, which forces the data to be processed in a series of overlapping tiles (done to prevent border artifacts). This network can likely be substantially accelerated, but we see acceleration as an orthogonal aim.

4.1. Recap: Training and Inference in SynthIA

The network is fit to minimize the divergence between the network's predictions and a target distribution. At a pixel with target value y (measured in Mx cm^{-2} for αB, for instance), we construct a target distribution $\boldsymbol{p}$ over the bin values $\boldsymbol{b}$ such that the expectation of the distribution $\boldsymbol{p}^{\top}\boldsymbol{b}$ is the target y. Specifically, we find the adjacent bins to y (or $\boldsymbol{b}_i \leqslant y < \boldsymbol{b}_{i+1}$), and then set $_{pi}$ and $_{pi+1}$ to make $\boldsymbol{p}^{\top}\boldsymbol{b} = y$. Then, if $\boldsymbol{z}$ is the log-likelihood of the bins produced by the network at a pixel, we minimize the Kullback–Liebler divergence (Kullback & Leibler 1951), which is minimized when the network's log-probabilities match those of the target distribution $\boldsymbol{p}$.

Following Higgins et al. (2021, 2022), we fit the network parameters (i.e., its convolutional filters and biases) by a stochastic gradient descent-like procedure. We identically follow Higgins et al. (2022) and train models using AdamW with a learning rate of 10^{-4}, $\epsilon = 10^{-4}$, and weight decay 3×10^{-7}, except for following a fixed learning rate for up to 100 epochs.

Once trained, one converts the network's distribution over discrete values into a final output. Here, we focus on a single pixel; the operation is applied identically to all the pixels in the image. Once the network is trained, the network produces the log-likelihoods of a sequence of bin values $\boldsymbol{b}$. These logits $\boldsymbol{z} \in \mathbb{R}^{K}$ can be converted to a distribution by applying a softmax operator σ: $\mathbb{R}^{K} \to [0, 1]^{K}$ that is defined as

$$\sigma(\boldsymbol{z})_i = \frac{\exp(\boldsymbol{z}_i)}{\sum_{j=1}^{K}\exp(\boldsymbol{z}_j)}, \tag{4}$$

which converts the logits into a distribution $\hat{\boldsymbol{p}} = \sigma(\boldsymbol{z})$ that sums to unity.

Given the distribution over bins, we decode the distribution to an estimated value. One option, which produces an intrinsically discrete distribution, is to then take the most

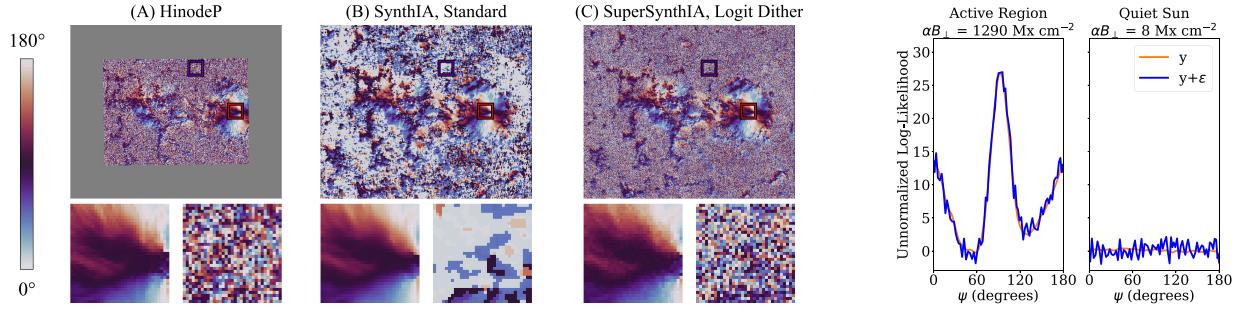

Figure 5. Mitigating unphysical preferences in azimuthal quantities with logit dithering. In the left panels, we show a cutout and then two close-ups centered on two regions. The left subpanel shows an active region patch where the azimuth is well defined with $\alpha B_\perp$ of 1290 Mx cm^{-2} at the central pixel. The right subpanel shows a region with poor polarization whose particularly weak central pixel has $\alpha B_\perp = 8$ Mx cm^{-2}. The azimuth should then be randomly distributed with no signal. On the right, we show the network's unnormalized log-likelihoods (logits) with and without logit dithering for nondisambiguated azimuth. HinodeP (A) is the result of an optimization-based Stokes inversion, and produces uniformly distributed azimuths in regions with low-polarization signal. Using the standard approach in SynthIA (B) of selecting the most likely bin results in preferential directions in the bottom close-up: the network slightly prefers an azimuth, and the preferred azimuth is repeatedly chosen. By adding logit dither (C), this preference is mitigated, and the output more closely resembles HinodeP. In the active region close-up, there is substantial signal, and the dither does not alter the prediction. Example: 2016 January 10, 19:24 TAI.

likely/modal bin m (i.e., $m = \arg\max_i \hat{\boldsymbol{p}}_i$) and provide its value $\boldsymbol{b}_\mathrm{m}$. Instead, Higgins et al. (2021, 2022) computes an expected value over the bins near the most likely bin. The value is the expectation over bin values adjacent to the modal bin m, or

$$v = \frac{\sum_{i=m-1}^{m+1} \hat{\boldsymbol{p}}_i \boldsymbol{b}_i}{\sum_{i=m-1}^{m+1} \hat{\boldsymbol{p}}_i} \tag{5}$$

where $\sum_{i=m-1}^{m+1} \hat{\boldsymbol{p}}_i$ renormalizes the distribution over the selected bin m and the adjacent bins.

One advantage over regression is that the network automatically produces more than a point estimate of the value: given a cumulative distribution $\hat{\boldsymbol{c}}_i = \sum_{j=1}^{i} \hat{\boldsymbol{p}}$ and desired quantile α, one can find the first bin i for which $\hat{\boldsymbol{c}}_i \geqslant \alpha$. In addition to finding the first bin surpassing the quantile, one can get a continuous estimate of the quantile's location by linearly interpolating. This procedure enables the estimation of the median (as the 50% quantile), as well as confidence intervals (by estimating, e.g., the location of the 5% and 95% quantiles).

4.2. Logit Dithering and Median Decoding

The above procedure produces unphysical artifacts that result from interactions between this decoding procedure and network training: network overconfidence leads to unphysical preferences in the predictions, and taking the most likely bin for many quantities results in oversmoothing of networks of plage. We now explain these issues and a mitigation strategy.

One unphysical artifact is preferred directions in azimuthal quantities in low-polarization regions—networks see structure even among noise. In an optimization-based inversion, regions with low-polarization signal are distributed approximately uniformly since *any* azimuth minimizes the χ^2 error between the synthesized and observed Stokes vectors. In Higgins et al. (2022) and other learning-based systems we trained, the deep networks developed preferential directions. This phenomena happens because deep networks replicate and often amplify biases in the data. These preferences can be in subsets of the data and do not have to be particularly strong. For instance, if the azimuth in low-polarization regions near the equator to the left of a sunspot is slightly more likely to be 110°, the network's outputs will reproduce this bias. Since decoding takes the most likely output, the slightly preferred value will be repeatedly selected, as shown in Figure 5.

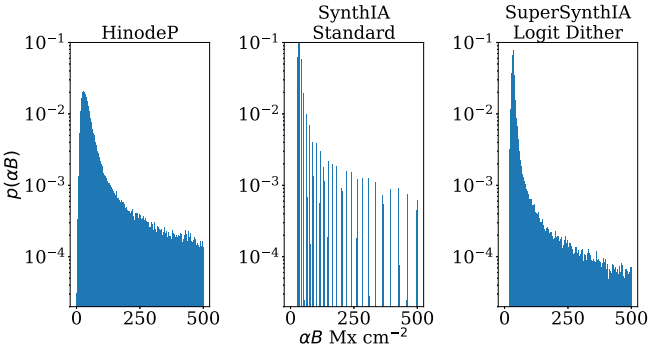

Figure 6. Banding artifacts: while banding is not typically visible in pictures of the outputs from SynthIA, it can be seen in the histograms (bottom row). We show histograms of αB from 0 to 500 Mx cm^{-2} with bins of 2 Mx cm^{-2}. While HinodeP shows a smooth distribution, the standard approach produces spikes due to network overconfidence, which effectively discretizes the predictions. Logit dithering produces smooth results that more closely match HinodeP. Example: 2016 January 10, 19:24 TAI.

A more subtle preference is a tendency toward predictions near the bin values. This preference is most clearly seen in histograms, where it appears as banding artifacts. While not usually visible in the outputs, they are unphysical and may lead to artifacts in the gradients of the outputs. These emerge from network overconfidence: deep networks tend to overestimate the probability of the most likely value, and so probabilities tend to be close to 0 or 1. Thus, while Equation (5) can produce any convex combination of the bins $\boldsymbol{b}$, the resulting outputs cluster near bin values. For instance, plotting histograms of αB values in Figure 6 reveals striping patterns in the standard approach as compared to a continuous distribution in HinodeP.

We mitigate both preferences with a procedure that is akin to dithering in signal processing. Rather than use the original logit $\boldsymbol{z} \in \mathbb{R}^K$, we sample a vector $\boldsymbol{\epsilon} \in \mathbb{R}^K$ from a distribution centered on zero with finite support and use a dithered logit $\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\epsilon}$. The distribution for the noise vector $\boldsymbol{\epsilon}$ is not large enough to change confident decisions, but nudges nonconfident ones. For instance, as shown in Figure 5, when estimating the azimuth of a strongly polarized pixel, the dither vector will not change the result; in a weakly polarized pixel with no signal, most of the final decision depends on the dither vector. Compared to sampling from the resulting distribution, dithering offers the advantage of not altering sufficiently confident predictions.

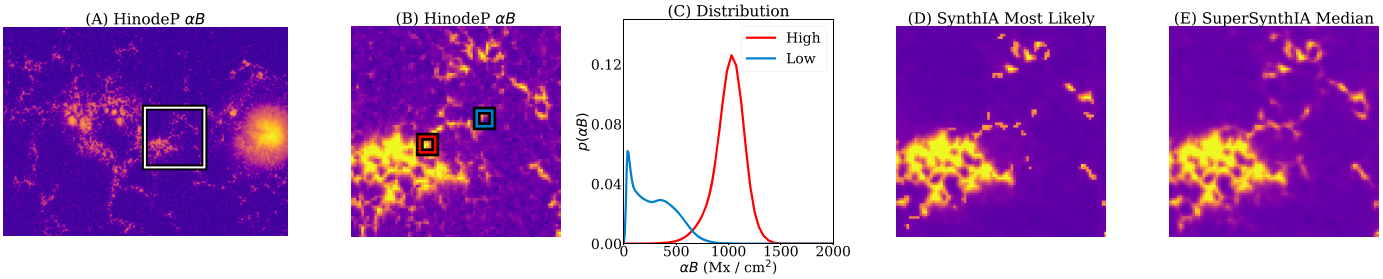

Figure 7. Distributions over αB inferred by deep learning method, and two methods to recover a single value. Left to right: (A) HinodeP αB for a full scan; (B) a zoom-in on the HinodeP αB; (C) the distribution inferred by SuperSynthIA, with a high value plotted in red and a lower value in blue; (D) the result from picking the most likely value as in SynthIA; and (E) the result from picking the median value, as done in SuperSynthIA. Note the substantially higher sharpness and sudden-drop-off in SynthIA as compared to the smoother fall-off in SupersynthIA. Best viewed on a screen in color. Colormap: 0 1000 Mx cm^{-2} for zoom-ins, and up to 3000 Mx cm^{-2} for the full cutout. Image example: 2016 January 10, 19:24 TAI.

In addition to banding and directional preferences, we found that networks tended to smooth out weak network regions to look like quiet regions as seen in Figure 7. The decision to pick the *most* likely value works well when the distribution is strongly concentrated around a single peak, but works poorly when there is substantial skew in the distribution. When there is skew, much of the distribution is concentrated away from the peak. These skewed distributions often happen in weak networks of plage. Instead of picking the most likely probability, we pick the median bin, which we define as the first bin for which the cumulative probability is above $\frac{1}{2}$. Using the median substantially pulls up the estimate of these weak regions, making the outputs look more like HinodeP.

Putting these improvements together in practice results in two variants of inference. Both use a dithered logit $\boldsymbol{z}' = \boldsymbol{z} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is drawn from a symmetric triangular distribution centered at zero with its scale a tunable parameter. We chose a triangular distribution due to its simplicity, concentration at zero, and finite support (which bounds how much decisions can change). We found it effective to sample $\boldsymbol{\epsilon}$ from a triangular distribution centered on 0 and extending from $-\sqrt{6}$ to $\sqrt{6}$ (which makes its variance unity). For azimuthal values (ψ and ψ_{360}), where we want to mitigate a directional preference, we pick the most likely bin and perform the expectation in Equation (5) using the dithered logits $\boldsymbol{z}'$. For other values, the most likely or median option for selecting the bin provides different trade-offs in terms of characteristics. Median decoding provides a smoother fall-off from active regions and plage into the quiet Sun, but results in a higher hemispheric bias as discussed in Section 6.3. Most likely decoding mitigates the hemispheric bias at the cost of sharpness. However, both result in quantitatively similar results, and switching between the two does not require retraining. In the paper, we report results using a median strategy: we select the median bin according to the original logits $\boldsymbol{z}$, and then compute Equation (5) using dithered logits $\boldsymbol{z}'$.

4.3. Fit Bins

As a last improvement, we fit the bins to the data rather than use regularly spaced ones as in Higgins et al. (2021, 2022). These bins are fit on a representative sample of values from the training set with k-means (MacQueen 1967) with two necessary modifications.

First, we include the minimum and maximum values of the value in question in the bin values, and fit $K-2$ clusters using K-means with $K-2$. During both inference and training, the output value is represented indirectly via a probability distribution $\boldsymbol{p} \in [0, 1]^K$, or $\boldsymbol{p}^\top \boldsymbol{b}$. This convex combination can only *interpolate* between bin values, and so, the extreme values are needed to cover all values.

Second, for targets that are symmetric about a target-specific value c, we generate bins that are also symmetric. If $d_{\max}$ is the maximum distance to c in the data set, we use bins $c - d_{\max}$, c, and $cd_{\max}$ with the rest determined by clustering the distances to c into $(K-3)/2$ clusters, and for each cluster d, including $c-d$ and $c+d$. We assume disambiguated vector components are symmetric about $c = 0$, and the inclination is symmetric about $c = 90$.

4.4. Direct-versus-analytic Heliographic Components

Finally, we present two paths to recovering the heliographic, physical magnetic field components $[\alpha B_R, \alpha B_\phi, \alpha B_\theta]$. The first is the *direct* method: estimating the heliographic (physical) components "directly," in other words performing the ME-like inversion and disambiguation simultaneously using the deep network, which has the advantage of being a single forward pass of a deep network per component. The second is *analytic* (or *"indirect"*): we estimate the image-plane αB, γ, and ψ_{360} (disambiguated azimuth) with the deep network, and then calculate the heliographic components analytically according to the pixel's location on the disk. On first glance, one would expect that both the direct and analytic methods would produce the same results, but in fact, they differ in their handling of the uncertainty and produce different artifacts.

In regions of weaker polarization signals, the network is often confident about the *length* of the vector (i.e., αB) but not certain about its orientation (i.e., γ, ψ_{360}). When the vector components are directly predicted independently in these weak-signal regions, the most likely solution is often setting all the components to zero, causing the total length of the field vector to be zero. In contrast, by predicting αB, and then orienting it via γ and ψ_{360} and the standard coordinate transform (Gary & Hagyard 1990), the vector length is nonzero. This effect leads to the analytic (indirect) method recovering data that more closely resemble HinodeP's overall field characteristics in weak polarization regions in poles, as shown by Figure 8. On the other hand, this analytical solution reduces on-average performance slightly, akin to how blurry images often better minimize mean-squared error objectives compared to more natural-looking sharper images.

We find differing artifacts in the outputs. The direct method tends to reproduce an artifact caused by errors in the pointing for HinodeP: the disambiguated components were calculated with the nominal Hinode/SOT-SP pointing information, which

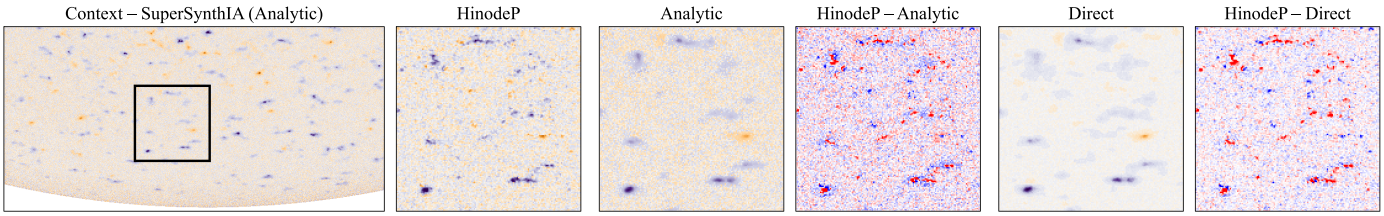

Figure 8. Comparison of two methods for obtaining αB_R from SuperSynthIA. Normally, the heliographic components are predicted directly by the model; in analytic prediction, the network predicts the in-plane components αB, γ, and ψ_{360}, and these are lifted to heliographic components following Gary & Hagyard (1990). Direct prediction of the vector components smooths out some of the data in weaker signal regions; analytic calculation results in recovery of more of these weaker regions. The result is that the approach of directly predicting αB_r has lower MAE, while the analytic approach more closely resembles the statistics of HinodeP. These differences are most pronounced at the poles. Colormap, -500 500 Mx cm^{-2}; difference colormap, -100 100 Mx cm^{-2}.

in turn leads to a sliver of the disk being always undefined due to the directional bias described by Fouhey et al. (2023). The direct method reproduces this bias and produces values of $\approx$0 Mx cm^{-2}, while the indirect method continues to produce reasonable estimates of field strength. At the poles, the quiet Sun direct estimates of B_ϕ and B_θ show a consistent unidirectional bias in quiet Sun, likely due to errors in pointing information that the network reproduces. In the disk center, in quiet regions, the direct method smooths out a lot of signal in B_ϕ and B_θ. Both of these blurred results are recovered by the indirect method. On the other hand, the indirect method has occasional azimuth errors inside sunspots: while the magnitude αB of the vector is estimated well, its orientation is not, leading to errors. We hypothesize that directly estimating the heliographic components is easier: the network has to only estimate a large αB_R, rather than an azimuth angle that varies by location.

Different downstream applications will have different purposes, and the two approaches offer different trade-offs. For equatorial cutouts, we report the direct method, and for polar cutouts, we report the analytic method. For full-disk images, we show a blended version. We create a per-pixel mask $\boldsymbol{M} \in \{0, 1\}^{H \times W}$, which is 1 where the analytic method is to be used. We set $\boldsymbol{M}_i$ to 1 in the top 200″, and a 32″ ring on the limb. To avoid a seam where the solutions switch, we create a soft mask $\boldsymbol{W} \in [0, 1]^{H \times W}$ that blends the two solutions. We set $\boldsymbol{W}_i = \exp(-d/\sigma)$ where d is the distance from the ith pixel to the nearest pixel in the mask $\boldsymbol{M}$ that is set to 1, and σ is an empirical smoothness parameter that we set to 150″.

5. Experiments

We now evaluate our new SuperSynthIA system, purely in terms of ability to estimate Hinode/SOT-SP-like vector magnetic fields. Our goal in this section is to explain the experimental setup, and understand the performance of the system on a per-pixel basis with its strengths and weaknesses in comparison to alternate approaches. We analyze artifacts, failure modes, biases, and temporal trends subsequently in Section 6.

In this section, HMIP serves as our primary *reference*, or baseline in the original sense in machine learning terminology. Our goal is not to determine whether HMIP or HinodeP or SuperSynthIA produces *better* magnetograms since no experiment in this section can do this for the remotely sensed data. Instead, the comparison to HMIP is critical to determining whether SuperSynthIA can produce magnetograms that closely resemble HinodeP. If one wants HinodeP-like magnetograms on the full-disk, the best current solution is HMIP since it produces excellent full-disk magnetograms. Due to its effectiveness, HMIP serves as a strong baseline. If the mean absolute error (MAE) between SuperSynthIA and HinodeP is larger than HMIP and HinodeP, then something likely has something wrong in terms of its data preparation or training. On the other hand, if a method produces better agreement with HinodeP compared to SDO/HMI, evidence would suggest that the method is extracting the information that lets it predict HinodeP-like magnetograms. Such a result is not a foregone conclusion and must be demonstrated experimentally.

5.1. Data Sets Used

We split the data set of cutouts into splits for training (for fitting parameters and model selection) and testing (evaluation) based on time. Our test set consists of the year 2016 to match Higgins et al. (2022), and build our remaining data outwards from this year, aiming to prevent data leakage, where the network has seen very similar data (e.g., the same active region) in the training set. First, we set aside two 6 month buffers (2015 July 1–2015 December 31 and 2017 January 1–2017 June 30) that do not appear in the training or validation set. These buffers correspond to 6.7 Carrington rotations. We use a brief window from late in the first buffer period for a consistency test. The remaining data (before 2015 June 30 and after 2017 July 1) are set aside for a training and validation.

We use four data sets for evaluating the method.

1. The first is the *cutouts* test set, which consists of all test cutouts from 2016 January 1 to 2016 December 31. The cutouts data set enables answering per-pixel accuracy questions (i.e., how well does a method agree with HinodeP?). Due to the narrow field of view of the cutouts, answering questions about systematic errors is more difficult.
2. The second is an 80 day data set, consisting of data sampled at an hourly cadence (starting at 36 minutes past the hour) for 80 days. To match Higgins et al. (2022), we pick 2016 February 1 as the start date. While HinodeP-like data are not available for the full-disk, this selection enables examining spatiotemporal artifacts on data that the network has not seen.
3. The third is high-cadence data sets covering the transit of two active regions (ARs) at 12 minute cadence—the highest-cadence publicly available Level-1 HMI Stokes data. In particular, we focus on NOAA AR 12422 (using 2015 September 23, 00:00 through 2015 October 3, 23:48) and 12567 (using 2016 July 15, 00:00 through July 22, 23:48).
4. The final is a decadal data set, consisting of data sampled at a 31 day cadence starting on 2011 January 1. This data set enables examining behavior over a long period of

time, albeit at the cost of clear leakage concerns. For each day targeted, we sample 12 observations on even hours at the 36 minute mark. The multiple samples enable testing at a variety of spacecraft velocities.

5.2. Metrics

To summarize per-pixel performance results, we use the same basic performance metrics introduced in Higgins et al. (2021): the MAE and the percent less than a threshold (%<t) (Scharstein & Szeliski 2002). We start with N ground-truth values $\{y_i\}_{i=1}^N$, which in this section are HinodeP data (although we stress that this is not to claim that HinodeP is the true solar magnetic field vector). Then, given N predictions $\{\hat{y}_i\}_{i=1}^N$, the MAE is defined as the MAE $\frac{1}{N}\sum_{i=1}^N|\hat{y}_i - y|$. The percent good pixels (PGP) is defined as the fraction of pixels that are correct to within an empirical threshold t, or $\frac{1}{N}\sum_{i=1}^N 1(|\hat{y}_i - y_i| < t)$. The threshold t is a problem-dependent empirical constant that defines a "good enough" prediction. For instance, Fouhey et al. (2013) set them to 30°, 22.°5, and 11.°25 for surface normal estimation on consumer photos. We set the thresholds following Higgins et al. (2022): 47 Mx cm^{-2} for B and αB, 5° for inclination, 7° for azimuth, and 0.1 for α. We set αB_R to 40 Mx cm^{-2} and set both αB_ϕ and αB_θ to 30 Mx cm^{-2} to account for the naturally smaller range. The PGP metric has the advantage of being more robust to a small number of strong outliers as compared to the MAE, but both metrics should be considered to paint a full picture.

To give a sense of overall overprediction/underprediction, we also compute the median signed error (ME), which is defined as the median of $\{y_i - \hat{y}_i\}_{i=1}^N$. This statistic gives a sense of overall bias in the predictions. Ideally, methods should have a median error of zero.

In addition, we plot bivariate histograms of the data that show the 2D joint distribution of y_i and $\hat{y}_i$. To aid with assessing bias and scatter, we also plot a red line showing $y = x$. For each prediction, we show a 67% empirical confidence interval in magenta and a 95% empirical confidence interval in white. The width of the interval gives a visual indication of scatter. If only one of the intervals comes close to the $y = x$ line, this indicates a strong bias.

We report results on three subsets of pixels. The first subset is *Equat*, or all the pixels that are aligned via the equatorial method. Since many equatorial pixels are from quiet regions, we additionally evaluate on equatorial pixels that have αB over 1000 Mx cm^{-2} as determined by HinodeP; we denote this subset *Equat1K*. This threshold limits the tests to include primarily well-determined pixels. Finally, *Pole* is all the pixels aligned by the polar alignment method. We suggest caution in interpreting the summary statistics at the pole and encourage consideration of the qualitative results: the long acquisition time leads to more remaining misalignment at the pole compared to more equatorial scans. Thus, the quantities are driven in large part by the quiet regions. Honing in on high field strength polar regions, however, may lead to incorrect estimates due to the misalignment. Since azimuthal quantities are poorly defined when the polarization signal is weak, we further filter these based on the strength of the transverse field using $\alpha B_\perp$. In all cases, we evaluate performance for ψ and ψ_{360} only where the HinodeP $\alpha B_\perp \geqslant 200$ Mx cm^{-2}.

5.3. Per-pixel ME Performance

We first verify the performance of SuperSynthIA on estimating ME inversion targets from HinodeP, primarily to verify the data preparation and machine learning. We note that the data preparation in this paper is different compared to the preparation done in Higgins et al. (2021). The change in data preparation makes direct table-to-table comparisons inadvisable, except as a rough approximation. Including HMIP also serves as a check on this data preparation: if HMIP's agreement with HinodeP is too low without a clear explanation, then there is likely something wrong with the data preparation.

We show and discuss qualitative results in Figure 9, and include a slice through Figure 9 in Figure 10. We also show bivariate histograms in Figure 11, including only αB at the pole in the bivariate histograms. As was the case with SynthIA (Higgins et al. 2022), SuperSynthIA shows better agreement with HinodeP compared to HMIP on equatorial scans. Up to ≈3750 Mx cm^{-2}, both SuperSynthIA and HMIP show good agreement with HinodeP. There is a pocket of high-flux pixels for which HinodeP provides a lower estimate than HMIP (3750 Mx cm^{-2} versus 4000–5000 Mx cm^{-2}). We hypothesize that this originates from the interpolation and rebinning to the SDO/HMI grid. An alternate explanation is the limited dynamic range of HMIP (Centeno et al. 2014) such that solutions from HMIP above ≈3200 Mx cm^{-2} are likely saturated; both HMIP and HinodeP invoke upper limits of 5000 Mx cm^{-2} for αB and B, respectively. It is difficult to see in the bivariate histogram, but HMIP estimates a substantially higher magnetic αB in quiet Sun regions compared to HinodeP. Instead, we encourage looking at two slices through αB in Figure 10 for the three methods: in addition to a higher quiet Sun value, HMIP also shows higher variation in quiet Sun regions.

With regards to the plane-of-sky inclination γ, HinodeP shows substantially more tilt to/from the observer as compared to HMIP (Sainz Dalda 2017), which is visible by the 66% intervals only touching the $y = x$ line. SuperSynthIA reproduces this behavior and, as shown in Figure 9, produces angles in plage that far better resemble Hinode/SOT-SP compared to HMIP, like its predecessor SynthIA (Leka et al. 2022). The discrepancy between the two pipelines' inclinations is discussed extensively in Sainz Dalda (2017), Leka et al. (2022). For the azimuth, SuperSynthIA has reduced scatter, visible by the lower probabilities away from the $y = x$ line. Compared to SynthIA, which showed strong preferred directions (as in Figure 5), SuperSynthIA produces roughly uniformly distributed angles in areas with weak polarization, as shown more extensively in Appendix C.

These results are further confirmed in Table 1: across the board, SuperSynthIA more closely resembles HinodeP as compared with HMIP. For pixels where HinodeP $\alpha B > 1000$ (Equat1000), HinodeP agrees nearly equally well with both SuperSynthIA and HMIP on αB. The good agreement in αB suggests that the overall performance on *all* equatorial pixels is driven mainly by differing nominal values for quiet Sun pixels. Indeed, the sign of the median error for HMIP on αB flips from Equat (a ≈50 Mx cm^{-2} overestimate) to Equat1K (a ≈43 Mx cm^{-2} underestimate). In turn, the relatively good agreement of HMIP and HinodeP on αB above 1000 Mx cm^{-2} suggests that differences in heliographic quantities are driven by the *angles* of the vector, rather than the magnitude: note that, in the Equat1K pixels, there is a 3.°4 MAE between

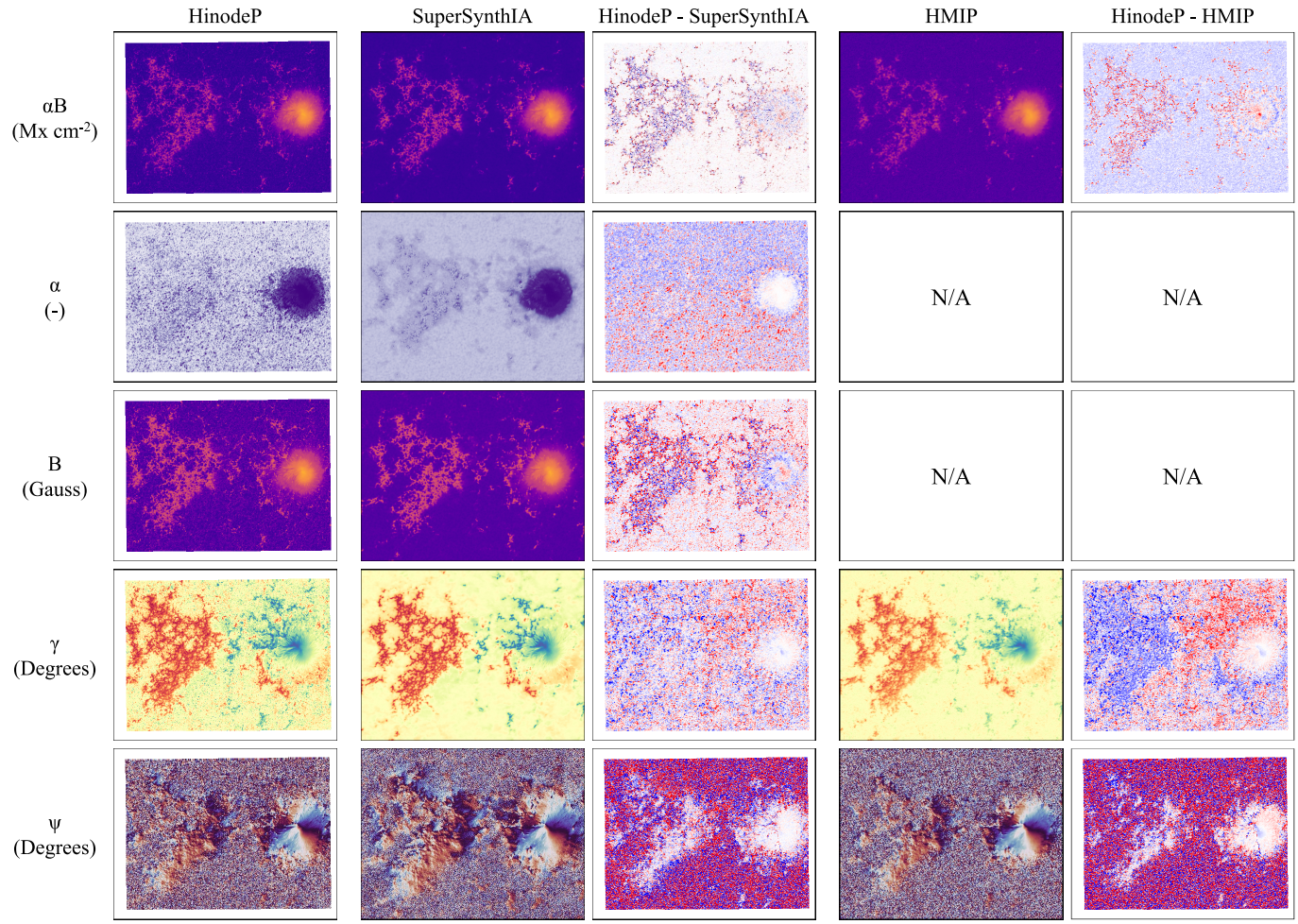

Figure 9. Qualitative results for ME inversion with HMIP shown as a reference. SuperSynthIA shows close agreement with HinodeP, and can separate αB into field strength B and fill factor α. In plage, SuperSynthIA produces inclination angles that better match the thick networks of highly inclined pixels seen in HinodeP. In azimuth ψ, SuperSynthIA infers a slightly larger coherent region of coherent field structure compared to either optimization-based inversion, but produces relatively unstructured field in areas with weak polarization. From top: αB (0 3000 Mx cm^{-2}; difference, -400 to 400 Mx cm^{-2}); fill factor α (0 1; difference, -0.5 to 0.5); field strength B (0 3000 G; difference, -400 to 400 G); inclination γ (0° 180°; difference, $-40°$ to 40°); azimuth ψ (0° 180°, difference $-40°$ to 40°). Difference map coloring: . Data: 2016 January 30, 16:00 TAI.

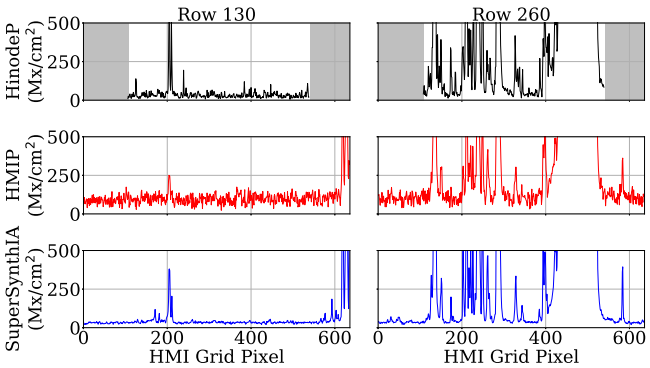

Figure 10. Two cross sections of αB from Figure 9 for (top to bottom) HinodeP, HMIP, and SuperSynthIA. For each method, we show row 130 (left) and 260 (right), roughly 1/4 of the scan down from the top row and the middle row of the scan. All methods correctly place the active region as seen in row 260, but SuperSynthIA shows less noise in quiet regions. As measured by the median absolute deviation in the row 130 cross section, HMIP has substantially greater variability across the quiet Sun row (21.6 Mx cm^{-2} compared to SuperSynthIA's 4.25 Mx cm^{-2}). Even after one normalizes by the 2.8× smaller quiet Sun value for αB from SuperSynthIA, this variability is much larger than HMIP. Dark gray regions in HinodeP are part of the buffer (see Figure 3) in which HinodeP data are not available.

SuperSynthIA and HinodeP in inclination, as compared to 6.8 for SDO/HMI and HinodeP. At the pole, HMIP substantially overestimates αB by nearly 100 Mx cm^{-2}. In particular, where HinodeP estimates $\alpha B < 50$, HMIP routinely infers $\alpha B > 100$. In contrast, SuperSynthIA produces substantially cleaner and quieter poles. This can be visually seen in Figure 12, where the background noise level is substantially higher in HMIP than in HinodeP or SuperSynthIA.

A table-to-table comparison with the SynthIA (Higgins et al. 2022) paper is tricky due to the improved alignment method presented in this paper. The seeming improvement over the results of SynthIA in B and αB is likely due to better alignment since the relative gap between SuperSynthIA and HMIP is similar. Azimuthal errors are lower because our evaluation is restricted to pixels with $\alpha B_{\perp} \geqslant 200$. To make direct comparisons easy, we directly compare in Section 5.6 to a SynthIA model.

5.4. Per-pixel "Direct" Heliographic Component Performance

We next analyze the performance of the method on direct estimates of heliographic components B_r, B_ϕ, B_θ. We show

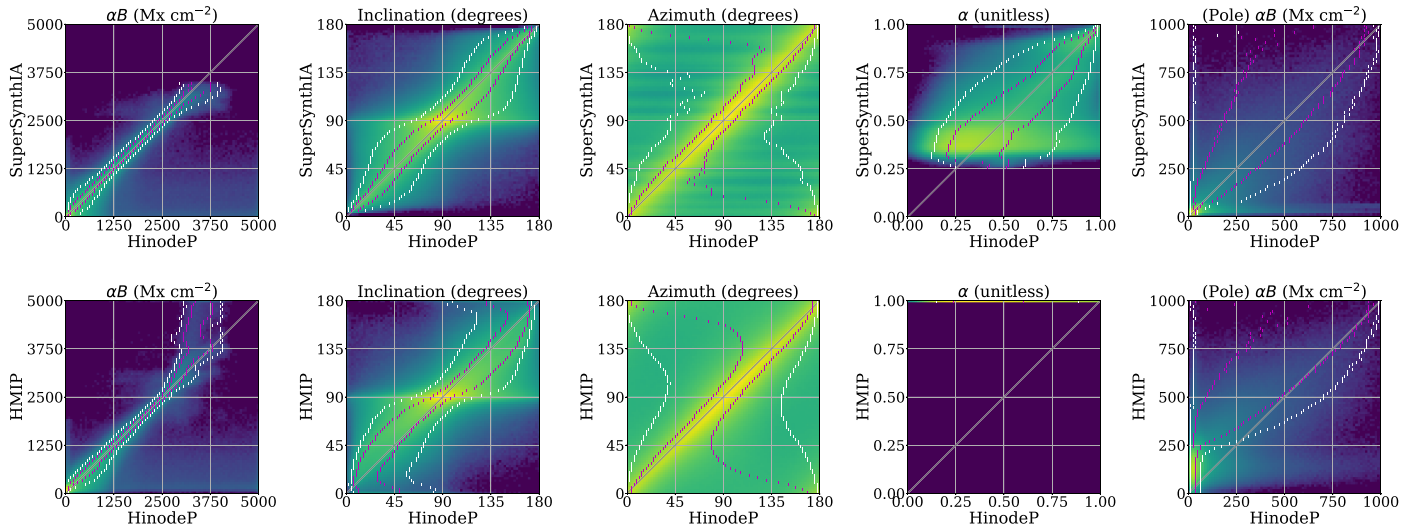

Figure 11. Bivariate log-histograms for Milne–Eddington (ME) Stokes inversion quantities. We histogram SuperSynthIA or HMIP against HinodeP, along with a $y = x$ line in dark gray, 67% confidence intervals in magenta, and 95% confidence intervals in white. The $y = x$ line touching or escaping the magenta intervals is an indication of substantial bias. HMIP inclination substantially deviates from HinodeP in ways that are understood: polar αB over all pixels is usually an overestimate due to the center-to-limb bias in low-signal pixels. Histogram legend: 0 Maximum log-density, where histograms have been normalized within a column to be directly comparable. Figure best viewed on a screen.

qualitative examples in Figure 13 and further results (including close zoom-ins) in Appendix C. As was done for the ME evaluation, we plot bivariate histograms in Figure 14 and summary statistics in Table 2. SuperSynthIA again has closer agreement to HinodeP as compared to HMIP. There are some pixels predicted by SuperSynthIA as having αB_ϕ or $\alpha B_\theta \geqslant 2000\,\mathrm{Mx\,cm^{-2}}$ while HinodeP predicts them as zero. We hypothesize that these are due to remaining misalignment since HMIP shows similar populations of pixels.

At the pole, HMIP and HinodeP substantially disagree on the value of B_r, as seen by the strong vertical bar for HinodeP $B_R = 0$. This disagreement originates from the similar disagreement on αB, which is visible in the plots of αB_R in Figure 12. In contrast, SuperSynthIA shows substantially less of such a disagreement.

5.5. Disambiguation Bit Estimation

We next further characterize the disambiguated azimuth. Recall that in the standard optimization approach of an ME inversion followed by a disambiguation, the disambiguated azimuth is the result of a 180° ambiguous azimuth and a disambiguation bit ($b \in \{0, 1\}$) via $\psi_{360} = \psi + 180b$. Directly testing this disambiguation bit agreement is not possible because SuperSynthIA directly maps to the disambiguated azimuth ψ_{360} without an explicit disambiguation. We therefore report acute-angle agreement, or the fraction of time that an estimated ψ_{360} lies in the half-circle centered on the reference ψ_{360}, or whether the vectors given by both angles have a positive dot product. We plot this agreement as a function of $\alpha B_\perp$ in Figure 15, using HinodeP as reference. Both HMIP and SuperSynthIA agree strongly with HinodeP in strong transverse fields, although SuperSynthIA produces better agreement in weaker fields.

5.6. Ablations and Comparisons

We conclude the pixel-by-pixel performance evaluation by comparing against two alternate approaches, regression and SynthIA (Higgins et al. 2022). The regression model is trained identically to SuperSynthIA but with a mean-squared error (i.e., $(\hat{y} - y)^2$) as opposed to the classification objective. SynthIA has the same backbone model as our method, but is trained on a smaller amount of data (a subset of 2012–2014) that is aligned with a more basic procedure, and its distribution is converted to a single value by taking the most likely bin without logit dithering.

We report quantitative results in Table 3. Our model usually outperforms regression and SynthIA. SynthIA's lower performance is likely due to its smaller training data set. Regression shows substantially worse performance on high field strength pixels (MAE 135.1 Mx cm^{-2} compared to 117.0 Mx cm^{-2}). We show why in Figure 16: models trained with a regression estimate tend to underestimate the true field strength.

The regression underestimation results from an interplay between the mean squared error (MSE) objective and residual misalignment in the data. For a target quantity y (here, e.g., αB) and a given probability distribution $p(y)$ over the value, the MSE is minimized by its expected value $\int y\,p(y)dy$. Suppose there is a pixel with a true value of 2000 Mx cm^{-2}, and 95% of the distribution is in a peak at the true value with some spread, with the remaining 5% placed ≈0 due to misalignment. The resulting MSE-minimizing prediction will be ≈1900 Mx cm^{-2}. Other regression losses *reduce* the influence of the outlier zeros, but typically do not eliminate it. In contrast, using regression-by-classification enables the model to account for the misalignment by placing some probability at 0.

6. Analyzing Artifacts and Trends

Having established that SuperSynthIA can estimate vector magnetic fields that are similar to HinodeP, we now turn to understanding artifacts and trends in the data. We see the analysis of artifacts and trends—both the good and the bad—as critical to the development of a learned magnetogram system.

In addition to examining failure modes, we focus on artifacts in space (e.g., center-to-limb biases), as well as in time (e.g., consistency or oscillations). This exercise serves several purposes. In some cases, SuperSynthIA improves on the existing HMIP data, and we seek to highlight these situations as potential

Table 1
Comparison of SuperSynthIA and HMIP with HinodeP on ME Quantities

		Field Strength (B)			Fill Factor (α)			(αB)			Inclination (γ)			Azimuth (ψ)		
		MAE	ME	$\% < t$	MAE	ME	$\% < t$	MAE	ME	$\% < t$	MAE	ME	$\% < t$	MAE	ME	$\% < t$
Equat	SuperSynthIA	87.4	5.3	51.8	0.1	−0.0	47.8	34.2	3.6	81.8	9.7	0.1	42.2	7.9	0.1	63.4
	HMIP	…	…	…	…	…	…	71.0	−52.7	37.4	11.2	0.3	36.2	9.6	2.4	57.4
Equat1K	SuperSynthIA	119.2	−11.1	35.1	0.0	0.0	90.0	109.7	15.3	39.9	3.4	0.1	82.9	5.5	0.0	79.4
	HMIP	…	…	…	…	…	…	138.8	43.6	35.7	6.8	0.8	54.6	7.1	2.4	71.8
Pole	SuperSynthIA	27.1	−7.7	88.1	0.1	0.0	76.5	11.1	0.0	98.2	8.8	0.0	42.8	14.5	0.5	41.1
	HMIP	…	…	…	…	…	…	100.7	−99.6	10.0	9.1	0.1	41.2	19.3	0.2	36.8

Note. We report how well SuperSynthIA and HMIP agree with HinodeP. HMIP's performance is a check to ensure that training was successful—SuperSynthIA ought to have better agreement with Hinode/SOT-SP compared to HMIP if trained properly. Azimuth results are only on pixels where $\alpha B_{\perp} > 200$ Mx cm^{-2}. We do not report α, B for HMIP since it does not give these quantities, although they are available via the method of Griñón-Marín et al. (2021). SuperSynthIA is closer to HinodeP compared to HMIP, indicating that SuperSynthIA is able to pull out the required information from the SDO/HMI spectra. Equat is nonpolar pixels; Equat1K is nonpolar pixels with HinodeP $\alpha B \geqslant 200$ Mx cm^{-2}; Pole is polar pixels.

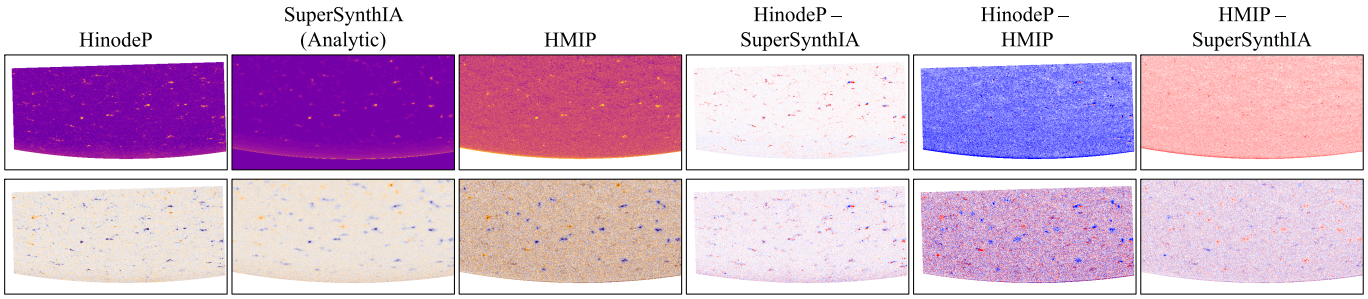

Figure 12. Polar estimates of αB (top) and αB_R (bottom) best seen on screen. SuperSynthIA produces reasonable estimates of polar αB and αB_R. The clear difference between HinodeP and HMIP is in low-signal regions that are usually masked with `conf_disambig`. There are patches of saturated colors in the difference maps between HinodeP and both HMIP and SuperSynthIA that do not appear in the difference map between HMIP and SuperSynthIA, and are due to residual misalignment: HMIP and SuperSynthIA place the bright spots at the same locations (as shown by the lack of large saturated areas in HMIP vs. SuperSynthIA difference maps) since they are interpreting the Stokes data similarly. Off-disk HinodeP data have been masked out using the SDO/HMI disk mask. Colormaps are αB 0 500 Mx cm^{-2}; αB_R −500 500 Mx cm^{-2}; and difference maps −200 200 Mx cm^{-2}. Data: 2016 September 1, 12:48 TAI.

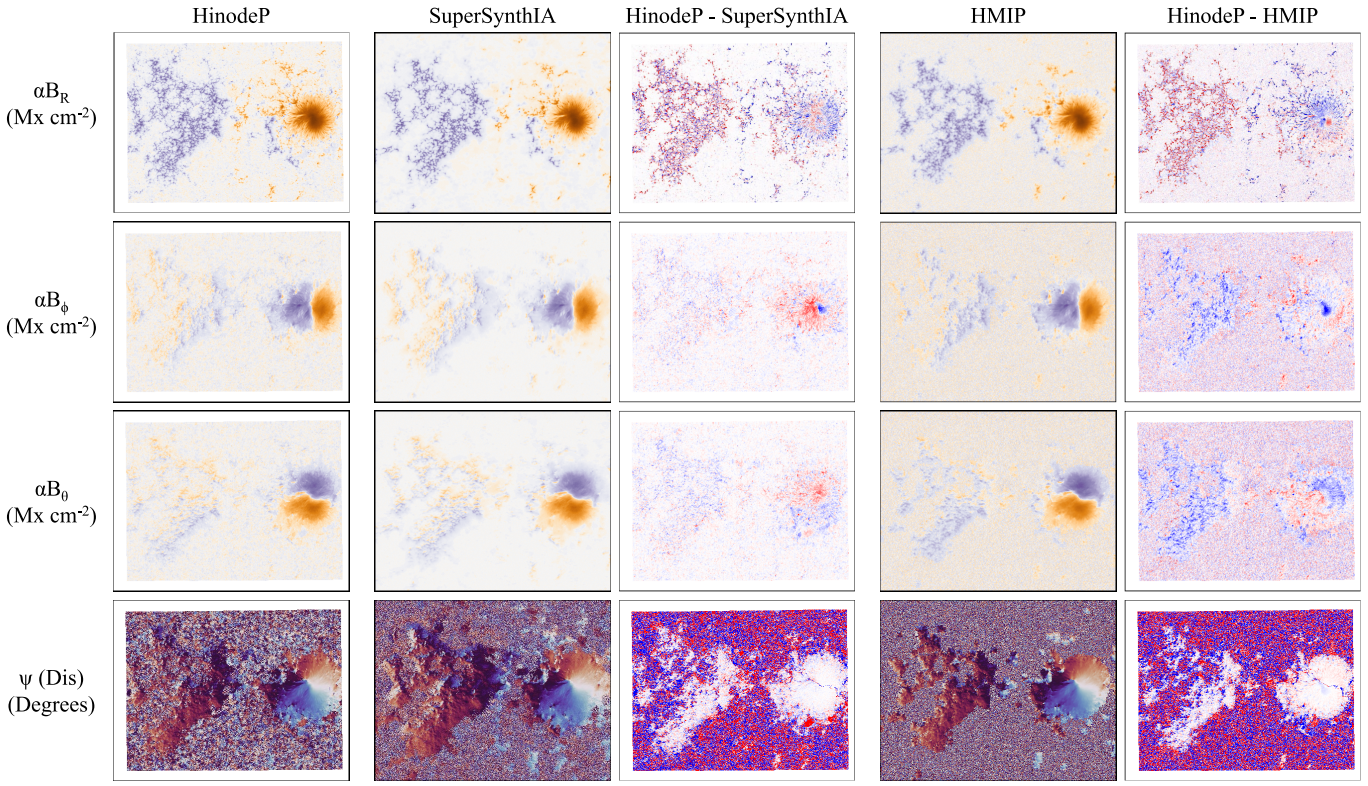

Figure 13. Qualitative results for heliographic quantities and disambiguated azimuth. SuperSynthIA shows fewer systematic deviations from HinodeP compared to HMIP. Colormap for heliographic components: (−3000 Mx cm^{-2} 3000 Mx cm^{-2}; difference, −400 400 Mx cm^{-2}); colormap for disambiguated azimuth ψ_{360}, (−180° 180°; difference, −80° 80°). Data: 2016 January 30, 16:00 TAI.

locations where downstream users may see specific benefit. In other cases, SuperSynthIA has limitations that end users need to be aware of: users may obtain poor or incorrect results or draw incorrect scientific inferences. Finally, in some cases, SuperSynthIA retains a trend or bias present in HMIP output. In the case of trends, this is a good result since HMIP is a trusted and well-calibrated source of magnetograms. In the case of biases, the reappearance of a bias in an entirely different inversion method can be illustrative about the data or the learning method.

6.1. Failure Modes

We start with known failure modes of SuperSynthIA. Since we hope that SuperSynthIA's outputs are used, we wish to make downstream users aware of potential issues. We looked for failure modes of the method by examining our 80 day data set as well as starting to prepare the outputs of SuperSynthIA for a number of downstream tasks. We report some below and visualize two of them in Figure 17 alongside five other examples for context. The system's current failure modes largely stem from the training data, such as errors in Hinode/SOT-SP's pointing and the asymmetric coverage that is created by having to upload targets in advance. These may be mitigated by more augmentation and improvements to various stages of data preparation.

Heliographic component results in large sunspots in poor coverage regions. A failure mode of two of our past papers Higgins et al. (2021, 2022) has been "collapses" in the centers of large sunspots, where extreme Zeeman splitting and very low intensity (photon counts) causes a collapse of information

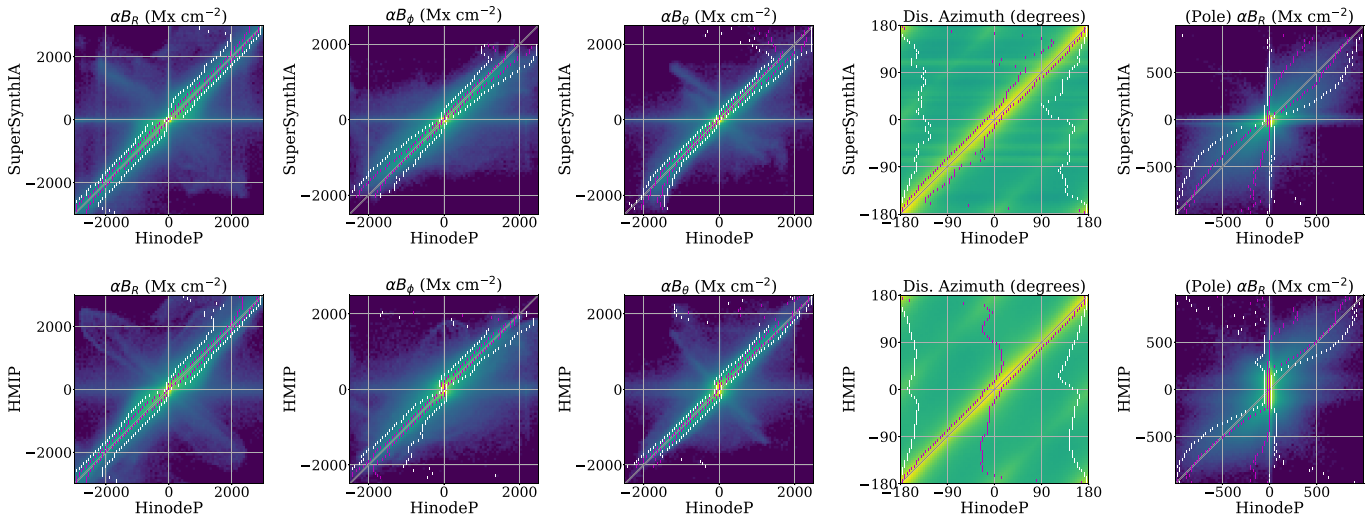

Figure 14. Bivariate log-histograms for heliographic quantities (αB_R, αB_ϕ, αB_θ) and disambiguated azimuth. We plot identically to Figure 11. SuperSynthIA and HMIP show good agreement in most cases. HMIP shows a barring effect best seen in αB_R at the pole: many pixels that HinodeP infers as near-zero Mx cm^{-2} are interpreted by HMIP as having upwards of a few hundred Mx cm^{-2} due to HMIP's noise pattern. These pixels are usually masked with `conf_disambig`. Colormap: 0 Maximum log-density. Histograms in a column share a colormap and can be compared. Figure best viewed on a screen.

Table 2
Results on Heliographic Quantities, Using HinodeP as Ground-truth and Reporting HMIP Agreement as a Reference

		αB_R			αB_ϕ			αB_θ			ψ_{360}		
		MAE	ME	$\% < t$	MAE	ME	$\% < t$	MAE	ME	$\% < t$	MAE	ME	$\% < t$
Equat	SuperSynthIA	33.1	−0.0	81.3	25.7	0.7	76.6	24.3	−0.3	74.4	11.4	0.0	82.9
	SuperSynthIA Analytic	34.8	−0.1	79.6	31.1	−2.1	65.8	32.7	0.9	57.5	11.4	0.0	82.9
	HMIP	54.6	−0.0	57.9	59.6	−0.1	30.2	67.3	−0.6	24.8	14.6	2.3	79.1
Equat1K	SuperSynthIA Direct	160.4	−2.4	22.1	150.7	20.6	17.9	97.9	4.7	24.8	8.4	−0.0	90.3
	SuperSynthIA Analytic	140.5	1.0	27.0	134.7	−7.2	23.6	94.8	−7.1	25.2	8.4	−0.0	90.3
	HMIP	181.9	−15.0	19.0	153.8	−3.3	18.1	116.4	−3.5	19.2	9.5	2.3	87.8
Pole	SuperSynthIA	13.7	0.0	95.6	13.6	−0.0	91.1	7.8	−0.0	98.3	25.6	0.1	61.2
	SuperSynthIA Analytic	22.8	3.7	86.3	23.1	−0.9	69.0	12.8	−0.8	93.2	25.6	0.1	61.2
	HMIP	78.0	1.3	26.7	80.0	−1.1	19.2	44.8	−1.3	38.5	42.5	5.5	52.5

Note. SuperSynthIA results are directly predicted αB_R, αB_ϕ, αB_θ; the heliographic components of SuperSynthIA Analytic are the result of lifting predictions of αB, γ, and ψ_{360} to heliographic components. Azimuth results are only on pixels with $\alpha B_\perp > 200$ Mx cm^{-2}.

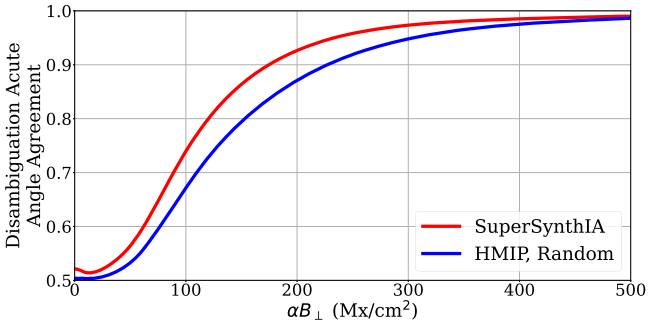

Figure 15. Acute-angle agreement between SuperSynthIA as a function of transverse field $\alpha B_\perp = \sin(\gamma)\alpha B$. We plot the rate at which the estimated ψ_{360} is within 90° of HinodeP ψ_{360}. Chance is 0.5, but for context, we also plot the HMIP acute-angle agreement, using the random weak-field disambiguation. Both have clear agreement in regions with strong transverse field; the remaining disagreement is likely due to residual misalignment. In weaker regions, SuperSynthIA has good agreement with HinodeP, slightly better than SDO/HMI, likely because SuperSynthIA is trained to emulate the ME0 method while SDO/HMI is using the random weak-field method.

Table 3
Comparison of SuperSynthIA with SynthIA and Regression on ME Quantities

		B	α	αB	γ	ψ
Equat	SuperSynthIA	87.4	0.12	34.2	9.7	7.9
	Regression	90.6	0.12	36.9	9.9	10.5
	SynthIA	97.5	0.13	38.0	10.0	7.5
Equat1K	SuperSynthIA	119.2	0.04	109.7	3.4	6.0
	Regression	128.7	0.05	201.6	6.0	9.0
	SynthIA	172.2	0.04	163.4	4.5	7.0
Pole	SuperSynthIA	27.1	0.07	11.1	8.8	14.5
	Regression	26.1	0.07	12.4	8.8	53.6
	SynthIA	49.1	0.11	22.8	9.0	12.1

Note. Azimuth results are only on pixels with $\alpha B_\perp \geqslant 200$ Mx cm^{-2}. SuperSynthIA usually does better than SynthIA or regression. Biases such as regression's systematic underestimate or SynthIA's preferential angles are not visible in this table, but can be seen in Figure 16 (for regression) and Figures 5 and 27 (for SynthIA).

in the spectropolarimetric signals. Occasionally, these extremely dark regions lead to "holes" where, amidst $\gg 1000$ Mx cm^{-2} fields, machine learning systems paradoxically estimate small areas of quiet Sun. These holes also occur in optimization-based outputs as noted by Hoeksema et al. (2014), and Hinode/SOT-SP has similar holes too. We observed far fewer of these in αB compared to our past works, likely due to

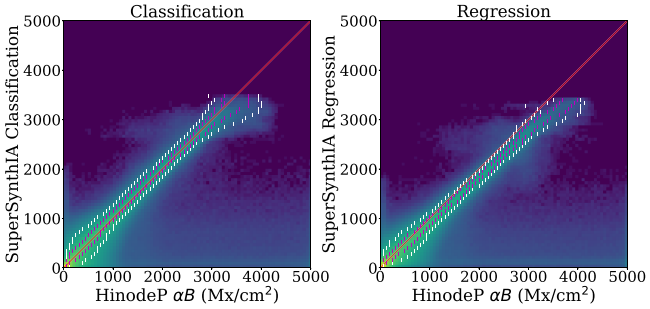

Figure 16. Regression vs. classification for αB. Regression systematically underestimates αB as seen by the $y = x$ line (plotted here in red) falling outside the confidence intervals. Classification more faithfully reproduces the HinodeP results. Figure best viewed on screen.

a larger data set and longer training times. In one case, we observed a failure in estimating heliographic components in an enormous sunspot; the failure occurred while the active region was on the limb with limited Hinode/SOT-SP training data, suggesting that data scarcity was at fault.

Reproduction of pointing issues in Hinode/SOT-SP. Hinode/SOT-SP is known to have inaccurate pointing information that has a strong bias in one direction (Fouhey et al. 2023). In addition to hampering disambiguation solutions, when the heliographic components are computed in the pipeline data, inaccurate pointing can lead to a ring of on-disk pixels being treated as if they were off disk. The current model that directly predicts heliographic components reproduces this ring, and thus has artifacts near the limb. These can be mitigated by using the analytic method for obtaining the heliographic components or by recomputing the HinodeP disambiguation components with corrected pointing and then retraining with the updated pointing information.

Bias in B_ϕ, B_θ at the poles. When making synoptic maps from SuperSynthIA, we discovered an asymmetric artifact in the polar quiet Sun in the direct predictions of SuperSynthIA: nearly all the quiet pixels point westward at one pole and eastward at the other. We hypothesize that this stems from the Hinode/SOT-SP pointing issues, which likewise show a directional bias. We believe that the network is amplifying a slight bias in the training data, and corrected data may mitigate these issues.

Oversmoothing. Our training data have been put onto the SDO/HMI grid, which is coarser than the Hinode/SOT-SP grid, and alignment requires a nonrigid correction to account for temporal evolution. While these steps are needed, these can erase small structures that are visible in HinodeP data.

Collapses in complex configurations. Inside some, but not most, complex magnetic field configurations during flux emergence, we observed occasional intermittent failures where regions were predicted as having low field strength.

6.2. Viewing Angle Dependence

The observer viewing angle should not make a difference in the inferred magnetic field since Earth's location does not change the magnetic field of the Sun. In other words, as the viewing angle changes from center ($\mu = 1$) to limb ($\mu = 0$), there should be no variation in the field strength. As explained in Hoeksema et al. (2014), in regions with a good polarization signal, HMIP produces good αB measurements in which there is no viewing angle dependence. On the other hand, in regions with a low-polarization signal, there is a known center-to-limb bias that users must mask out with a confidence mask (`conf_disambig`).

As shown in Figure 18, when evaluated on 500 scans from the 80 day data set, neither HMIP nor SuperSynthIA show viewing angle dependence in active regions. Following Higgins et al. (2021), we define these active regions as pixels for which `conf_disambig` $\geqslant 60$ and `Ic_noLimbDark` $\leqslant 0.8$. SuperSynthIA shows a slight dip in extreme viewing angles with 15° of the limb. However, apart from some natural variation due to solar evolution and limited sampling, αB is largely constant from 15° inwards. This constant value confirms that both HMIP and SuperSynthIA should be appropriate for analyzing active regions.

On the other hand, when one considers *all* pixels and includes low-signal pixels, HMIP shows a monotonic increase as a function of μ while SuperSynthIA largely does not. HMIP's increase happens steadily across all μ values and is accompanied as well by an increase in the variance. On the other hand, SuperSynthIA only starts diverging from its near constant αB at around $\mu = 0.25$ or $\theta = 75°$ and only diverges to $\approx 100\,\mathrm{Mx\,cm^{-2}}$. Results in Appendix A verify that this trend continues when one evaluates the error with respect to HinodeP.

The reader may be curious about the choice of active regions as opposed to simply picking confident pixels. This is because the distribution over αB for confident pixels itself changes as a function of μ. First, thresholds in the confidence mask change as a function of μ to account for the μ-varying properties such as noise. Second, as active regions transit the disk, their apparent size changes with μ due to foreshortening, which shifts the distribution. In combination, there are two independent μ-dependencies that cannot easily be accounted for. By selecting only active regions, there should be no μ-dependency: while there are *more* pixels due to foreshortening, the underlying distribution of flux remains steady.

Together, these suggest that SuperSynthIA may be usable toward as far as $\mu = 0.25$, covering nearly all of the pixels on the disk. The additional steadiness in low-signal regions suggests that users may be able to directly use SuperSynthIA's results for full-disk analysis. Users are currently able to filter out the noisy pixels in HMIP, but SuperSynthIA's natural noise reduction may offer benefits, for instance, in analyses where masking out pixels is difficult algorithmically.

6.3. Hemispheric Bias in Inclination, B_ϕ, B_θ

We next analyze the hemispheric bias discussed at length by Pevtsov et al. (2021), Liu et al. (2022), Griñón-Marín et al. (2021), Leka et al. (2022), and others. In short, an overestimation of $B_\perp$ in ME inversions can result primarily due to an erroneous handling of the inferred magnetic fill factor (here α), plus incomplete correction for scattered light, weighting factors, etc. The manifestation is clearest as a bias in direction according to position on the disk, and most obviously for B_ϕ as a target transits the central meridian. For a single number summarizing performance, we report the Gini-coefficient-based metric described by Leka et al. (2022) that ranges from 0 to 1, where 0 is no bias, and lower is better. Since the image-plane inclination angle is a primary driving force of this bias, we conduct an analysis where αB estimates and azimuth angles are held constant.

HinodeP itself does not show this bias (Leka et al. 2022), but the *sampling* of scans differs between the east and west

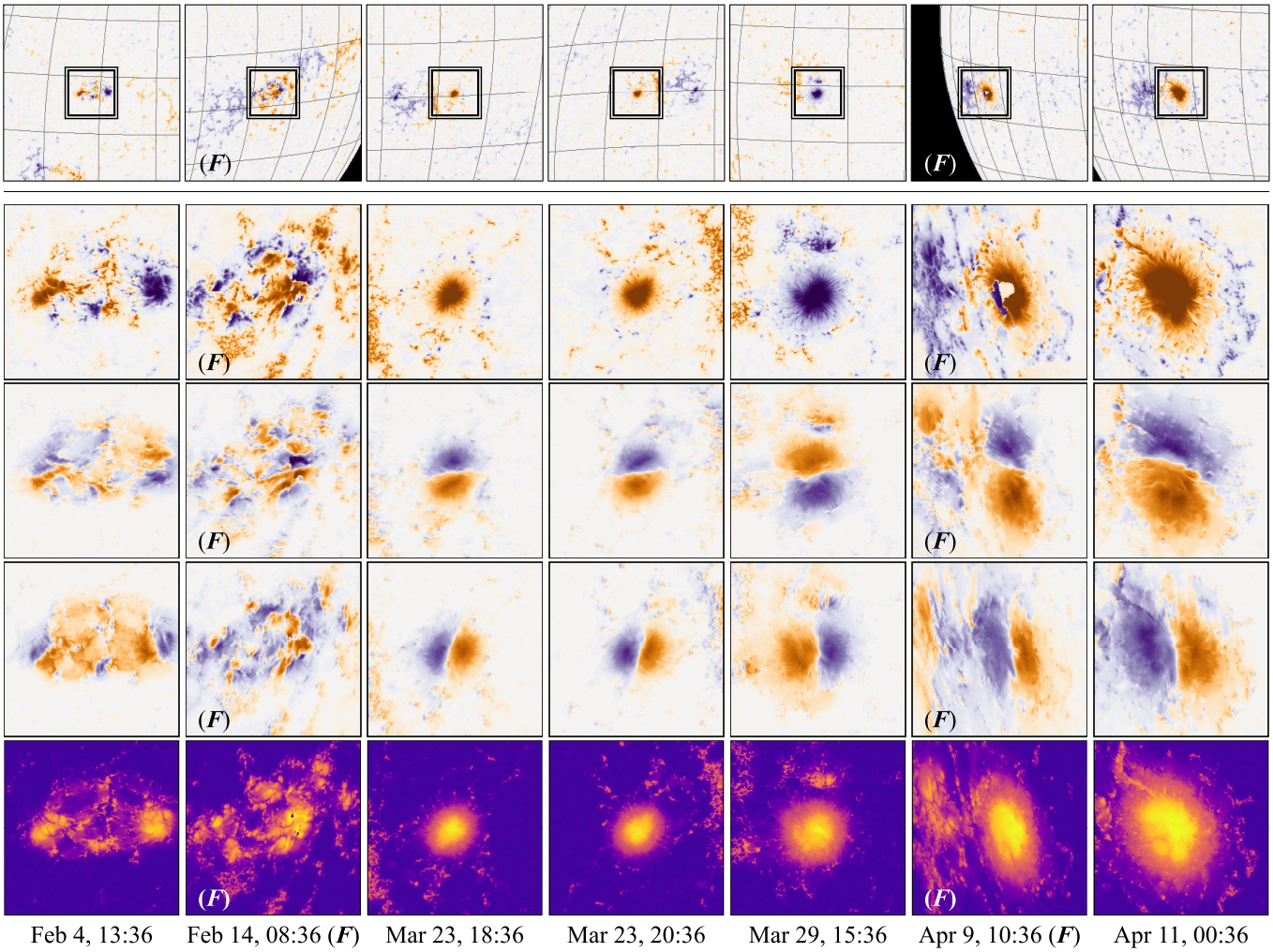

Figure 17. Examples from the 80 day test set (2016 Feb 1–Apr 20), with two failures marked with (*F*) in the bottom left of each image in the column. Top to bottom: αB_R giving the context of the cutout with 10° latitude and longitude lines for reference, and then 128″ cutouts for αB_R, αB_θ, αB_ϕ, and αB. Left to right: Feb 4 and 14, SuperSynthIA generally does a good job of interpreting complex configuration, although we *occasionally* observed small holes in strong-signal areas among particularly complex regions; performance like Feb 4 is far more frequent, and most misinterpretations are near the limb. Mar 23, a calibration maneuver that rolled SDO by 180° provided a natural consistency check. SuperSynthIA's interpretation of the sunspot is consistent when flipped. Mar 29, a typical sunspot. Apr 9, briefly, SuperSynthIA misinterprets the radial component αB_R (but not αB) of a sunspot that had just rotated on disk; this region has poor Hinode/SOT-SP coverage, and such large regions are fairly rare (as are these failures). Similar issues can impact optimization-based inversions for both SDO/HMI and Hinode/SOT-SP. Apr 11, once the sunspot moves farther onto the disk, the outputs have cleared up. Colormaps: −2000 2000 Mx cm^{-2}, 0 5000 Mx cm^{-2}.

hemisphere, stemming from the selection of Hinode/SOT-SP targets, sometimes 2 days in advance. Figure 4 shows the difference in the frequency of scans between east/west, but we also observed a difference in distribution of vector fields. We found that machine learning models trained with both regression and classification losses exploited this difference in the distribution across hemispheres as a statistical shortcut (Jabbour et al. 2020; DeGrave et al. 2021). In other words, simply identifying east versus west provided useful information about the probable distribution of inclination angles. With increased training, models relied on the shortcut less and focused on the signal present in the Stokes vectors as opposed to the simpler shortcut. This could be observed via a consistent decrease in Gini coefficient across regression and classification models as a function of training time. However, short of training for an infinite time, this shortcut remained present in the models.

Different approaches produced different amounts of bias. We obtained the best results, a Gini coefficient of 0.15, when the most likely bin was selected, following the SynthIA method and what was done for the azimuthal angle. The bias is slightly lower than HMIP, and we hypothesize that continued training will reduce this further (since the Gini coefficient decreased by about ≈0.03 for each of the proceeding 25 epochs, or passes through the training set). Using the median approach resulted in a Gini coefficient of 0.31 (which also was decreasing at a rate of ≈0.03 per 25 epochs), higher than HMIP. The median's higher bias is likely because the use of the statistical shortcut can adjust the tails and therefore change the median. Finally, models trained with regression produced the largest bias (0.35), with a slower rate of decrease in Gini coefficient over training. This is because the regression model produces the expected value, which is far more easily changed by events in the tail (i.e., driven by the statistical shortcut).

In applications where the hemispheric bias is likely to be a concern, users should prefer the most likely decoding for inclination and B_ϕ and B_θ. This produces a similar performance to the median decoding in overall metrics.

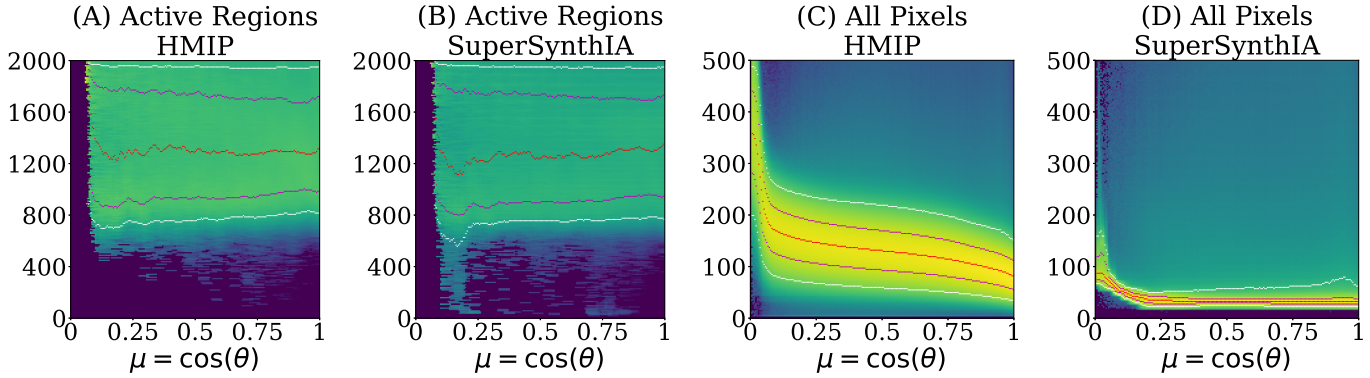

Figure 18. Distribution of αB as a function of disk location/viewing angle μ for active region pixels (left) and all pixels (right). On top of the log-density (minimum maximum log-density), we plot the median in red, 66% confidence intervals in magenta, and 95% confidence intervals in white. In principle, there should be *no* relationship between viewing angle μ and the magnetic field strength, and αB ought to be constant apart from any evolution in the data. Both HMIP (panel (A)) and SuperSynthIA (panel (B)) are steady for active region pixels and respect this relationship, albeit with a slight dip at extreme viewing angles for SuperSynthIA ($\mu \approx 0.15/\theta \approx 85°$). On the other hand, there is a well-known HMI artifact (Hoeksema et al. 2014) where low-signal regions show a steady increase in αB as a function of viewing angle (panel (C)). These regions need to be removed by a confidence mask, but SuperSynthIA is able to reduce noise in these regions, producing more consistent results in (panel (D)).

6.4. Vector Field Consistency over Minutes

We next turn to temporal artifacts. High-cadence photospheric vector magnetogram data ought to enable the physical interpretation of solar evolution, and the direct driving of MHD models. Any temporal inconsistencies that arise from artifacts impact downstream applications by suggesting unrealistically fast reconfigurations of the field. Here, we analyze the temporal consistency of SuperSynthIA and HMIP at a 12 minute cadence, the standard pipeline cadence available for SDO/HMI data.

We compare two approaches using SuperSynthIA to obtain temporally consistent time series data. The first is simply SuperSynthIA "direct," which effectively solves the inversion and disambiguation simultaneously and provides the heliographic components directly. The second is (*HMIP SS-Acute*) which uses HMIP's inversion output, but applies an acute-angle disambiguation to this output based on the results of SuperSynthIA. In particular, given the SuperSynthIA magnetic field vector $\boldsymbol{B}_{\rm SS}$ in heliographic coordinates, we compute the two possible vectors ($\boldsymbol{B}_1$ and $\boldsymbol{B}_2$) given by HMIP ME inversion followed by transformation to heliographic coordinates. We then select whichever of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ forms a larger dot product with $\boldsymbol{B}_{\rm SS}$. HMIP SS-Acute does not gain any artifact-reduction benefits of SuperSynthIA with regards to αB and direction, as it effectively replaces just the Minimum Energy disambiguation step of HMIP. However, by retaining the estimates of αB of HMIP, it therefore may be useful for applications that have been tuned carefully for the SDO/HMI pipeline.

We compare these two approaches to HMIP results, on two active regions. The first is NOAA AR 12567 from 2016 July 15, during the build up to the largest solar flare of 2016 (an M7.6 flare). The second is NOAA AR 12422, from 2015 September 25. Side-by-side video examination of both SuperSynthIA and HMIP SS-Acute against HMIP suggests that both of the former provide far more temporally stable output. The stability is best seen in a video, but we give an intuition by showing cutouts from AR 12567 in Figure 19. HMIP is often stable in sunspots, but can present unstable results, especially in regions that are poorly modeled by the ME assumption for which the Minimum Energy disambiguation cannot then find an appropriate solution. In contrast, both SuperSynthIA-based results provide temporal stability although the HMIP SS-Acute output lacks other properties, such as the sharpness provided by SuperSynthIA.

We quantify the results by examining the observation-to-observation consistency. Given field vectors at temporally adjacent observations $\boldsymbol{B}_{t-1}$, $\boldsymbol{B}_t$, and $\boldsymbol{B}_{t+1}$, we use two measures of consistency. *Distance* simply measures the difference between two adjacent frames, or the standard Euclidean distance $||\boldsymbol{B}_t - \boldsymbol{B}_{t+1}||$. Distance can be caused by both flickering inconsistencies, where an interpretation changes rapidly back and forth, as well as evolution, where a pixel changes value due to a region expanding or contracting. We therefore also define *Flicker Distance* as $\min(||\boldsymbol{B}_{t-1} - \boldsymbol{B}_t||, ||\boldsymbol{B}_t - \boldsymbol{B}_{t+1}||)$, which is large if and only if the vector at time t is far from both the previous and next observations (i.e., the field suddenly changes then reverts).

We evaluate both distance and flicker distance on a cutout following the active region. We start by linearly interpolating between waypoints marked each day on the active region, followed by frame-to-frame realignment of a $150'' \times 150''$ cutout to minimize the mean-square distance between the vector field and the previous frame's cutout. While aligning, we search over 10 pixels of displacement (in whole pixels). When computing consistency metrics, we repeat this procedure, searching up to two pixels. To ensure that results are as fair as possible to HMIP, the original initial cutout alignment procedure is done using HMIP vector fields, and we further restrict evaluation statistics to confident pixels (`conf_disambig` $\geqslant 90$). The use of only confident pixels in turn restricts the analysis to pixels where the ME0 disambiguation algorithm was run, and so, which fast approximation was used in the weak field makes no difference.

As shown in Figure 20, SuperSynthIA provides a clear reduction in flickers and inconsistency for large distance thresholds. For AR 12567, HMIP shows 60× as many flickers of over 1000 Mx cm^{-2}, and 80× as many flickers of over 2000 Mx cm^{-2}, as compared to SuperSynthIA. Even HMIP SS-Acute produces strong gains: HMIP has 3.4× as many flickers over 1000 Mx cm^{-2} and 10.6× as many over 2000 Mx cm^{-2}. The apparent slight potential improvement in consistency by HMIP at $<$500 Mx cm^{-2} in some cases is an artifact of evaluating only on pixels those pixels for which HMIP is confident.

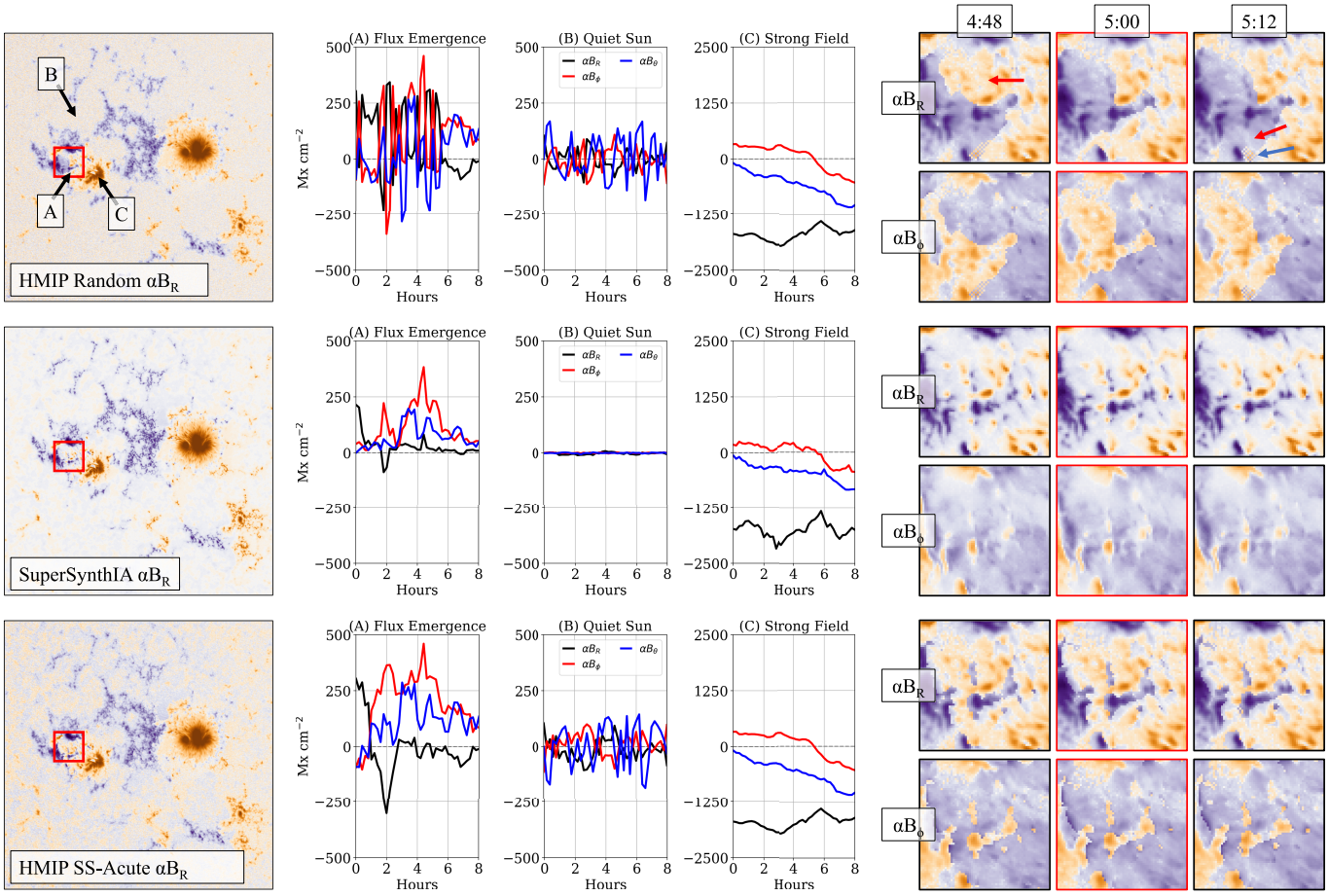

Figure 19. Short-term consistency for AR 12567 across time and space during the buildup to the largest flare of 2016. We show HMIP (top), SuperSynthIA (middle row), and HMIP SS-Acute, or using SuperSynthIA to pick the disambiguation bit (bottom). (Left): we show αB_R for each method in a $150'' \times 150''$ cutout at 2016 July 15, 5:00 TAI. (Middle): we plot αB_R, αB_ϕ, and αB_θ as a function of time at three pixels (in a region of flux emergence, in quiet Sun, and in a strong-field region). All methods show relatively constant evolution in strong-field regions. At the flux emergence pixel, HMIP shows rapid reversals as a function of time while both SuperSynthIA and HMIP SS-Acute evolve smoothly. At low-signal quiet Sun pixels (often suppressed by the `conf_disambig` mask), HMIP shows rapid reversals, while SuperSynthIA is smooth at near zero. HMIP SS-Acute shows these reversals, since it cannot alter the *length* of the vector. (Right): we show three $32'' \times 32''$ zoom-ins in the region of flux emergence at 4:48, 5:00, and 5:12 TAI. In the top plot, HMIP shows sudden magnetic field reversals, indicated with red arrows as well as occasional checkerboarding indicated with the blue arrow. On the other hand, SuperSynthIA shows more consistent estimates and no checkerboarding. This consistency is best seen in the accompanying animation. This animation is of the data from the middle image (SuperSynthIA) and top image (HMIP) from the first column of the figure. In the accompanying animation, the full solar disk is displayed on the left next to images of zoom-ins on the active region. SuperSynthIA is shown in a red box, and the HMI Pipeline is shown in a blue box. The animation begins at 2016 July 15, 5:00 TAI and is shown at 12 minute cadence. Colormap: -2000 2000 Mx cm^{-2}.

(An animation of this figure is available.)

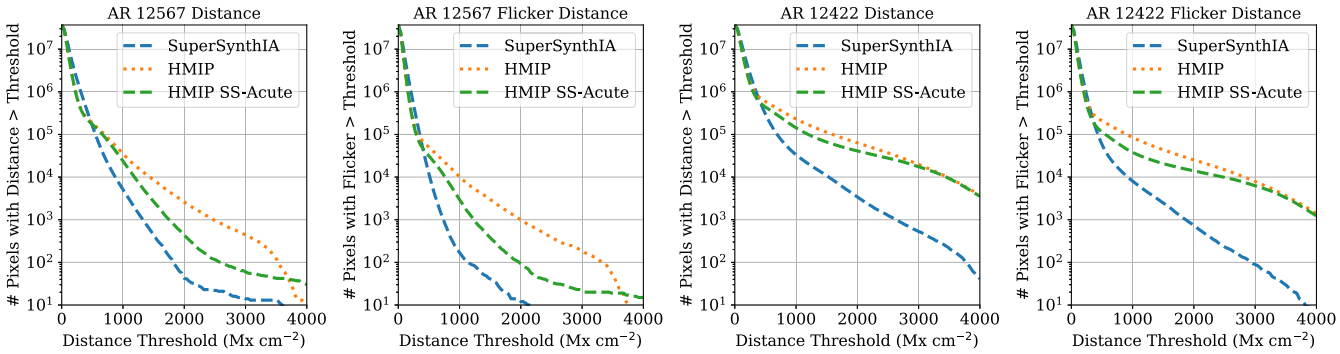

Figure 20. Plots of number of changes above a threshold as a function of threshold. SuperSynthIA reduces number of changes by an order of magnitude for most large thresholds if not more. We evaluate three methods, including SuperSynthIA, HMIP, and HMIP SS-Acute (picking the HMIP disambiguation bit to be closest to SuperSynthIA) and use two metrics, a standard distance and one only for flickers. All results are frame-to-frame distance after optimally registering the frames to minimize the mean-square distance, and results are evaluated only on pixels where HMIP disambiguation confidence (`conf_disambig`) is $\geqslant$90 for all pixels involved. Using *only* the pixels that HMIP is confident about biases the results toward HMIP, and is the root cause of its apparent comparable-or-better performance for lower thresholds like $\approx$500 Mx cm^{-2}. As shown in Appendix B, SuperSynthIA-based methods outperform HMIP at all thresholds when all pixels are used.

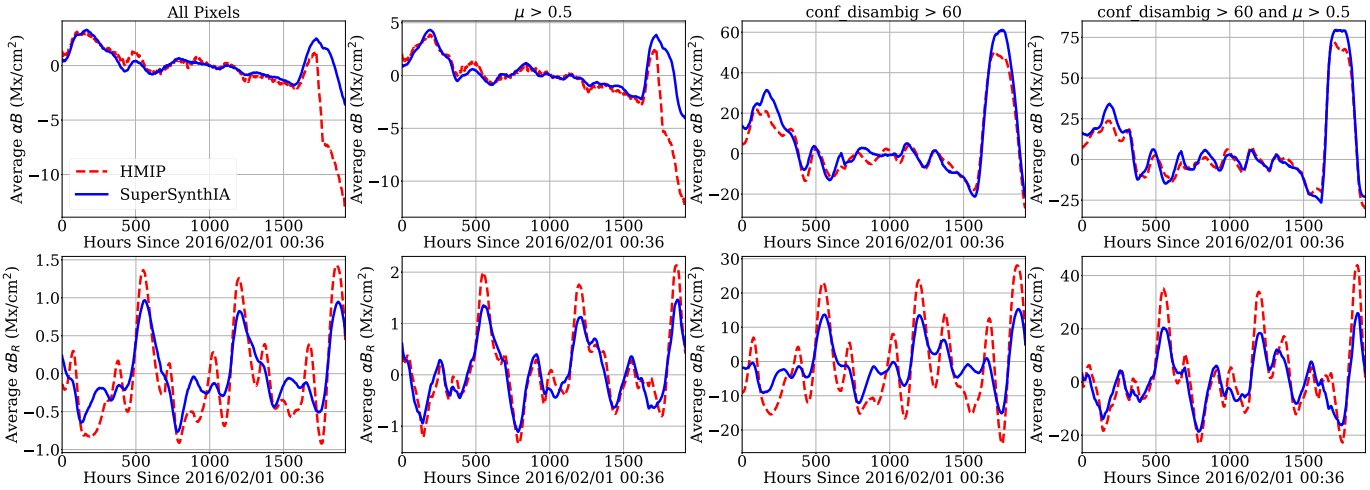

Figure 21. Around the Sun for 80 days. We plot the average αB (after subtracting the median) and αB_R for 80 days for multiple subsets of pixels: all pixels, disk-center pixels ($\mu > 0.5$), pixels with confident disambiguation (`conf_disambig` $\geqslant 60$) and confident disk-center pixels. The αB data are reported relative to the median to account for the difference in nominal quiet Sun values between HMIP and SuperSynthIA. To show the long-term trend, rather than short-term oscillations, we plot the average of a 48 hr window. In all cases, the SuperSynthIA output closely follows the HMIP output, for both αB_R (where it matches without adjustments) and αB (where it matches closely after subtracting the median).

Many of the regions in which SuperSynthIA provides more temporally stable output are centered on phenomena whose emergent Stokes vectors may challenge the SDO/HMI pipeline methodologies. Highly dynamic and unresolved fields can produce Stokes spectra that are inconsistent with the ME assumptions (for example, Sigwarth et al. 1999; Franz et al. 2016, and others). Nonunique inversion solutions are thus available over a multipixel area for multiple time steps, i.e., the inversion is unstable; furthermore, when an unstable or unphysical inversion output is fed to the disambiguation module, the results are similarly unstable.

We hypothesize that the stability of SuperSynthIA can be attributed to both the method and the source data. The most important factor is that SuperSynthIA is a finite capacity model that has been optimized for many pixels simultaneously across the training set. So, while it is possible that the SDO/HMI Stokes spectra are underconstrained by the ME model, resulting in multiple possible solutions, SuperSynthIA tends to pick one solution. A second likely factor is that the considerably more constrained inversion problem solved by HinodeP produces a more stable solution; while also an ME inversion, HinodeP utilizes higher-fidelity spectra, higher spatial resolution (natively), and many more spectral samples including a second spectral line with a different Zeeman and thermodynamic response, such that the HinodeP ME model can provide a more consistent minimum. In turn, this minimum can be identified by SuperSynthIA through the use of multiple SDO/HMI pixels.

6.5. Vector Fields over Days

The HMIP suffers from nonphysical variations in the magnitudes of the image-plane components, dominated by a 24 hr periodicity but with harmonics (Hoeksema et al. 2014; Schuck et al. 2016). This effect is instrument related and can be seen in Figure 5 of the Hoeksema et al. (2014) paper discussing the vector field of HMIP and Figure 9 of our previous work introducing the base machine learning model (Higgins et al. 2021). The HinodeP data appear to be free from this artifact. A careful analysis of the underlying data in the 80 day series indicates that SuperSynthIA (as well as its predecessor, SynthIA) unfortunately does not fully mitigate these oscillations.

If one examines the average of *all* on-disk pixels, SuperSynthIA shows a *substantial* $\approx 4\times$ reduction in the oscillation, just as in SynthIA. However, when the average is restricted to *only* high-polarization-signal pixels (`conf_disambig` $\geqslant 60$), there is little-to-no reduction in the oscillation. This seeming discrepancy is explained by a strong difference in nominal values for low-signal quiet Sun values. As shown in Figure 18, the SuperSynthIA quiet Sun data are far lower in inferred field strength and far less noisy compared to HMIP. Since quiet or low-signal pixels dominate any full-disk average, the smaller overall signal manifests as having a smaller oscillation magnitude as compared to the same averaging performed for HMIP output with its greater weak-signal noise and hence oscillation magnitude. While some downstream applications may still benefit from this overall noise reduction, the continued presence of this artifact in high-signal pixels indicates that its ultimate source is likely "upstream" of the inversion and impacts the input polarization directly. While more carefully selected training data and inputs may help mitigate this artifact, and there may yet be an ML-based solution, SuperSynthIA does not, by itself, fully mitigate the orbital-related oscillations.

6.6. Vector Fields over Weeks

While the results so far have established that SuperSynthIA more closely resembles HinodeP and is more temporally consistent between neighboring magnetograms, many modeling applications are critically sensitive to accuracy in the longer-term trends in the magnetic field configuration.

We examined the average αB and the average signed radial field αB_R over an 80 day period starting on 2016 February 1, employing different pixel selections to include (Figure 21). An average is taken over a 48 hr window to remove the influence of the orbit-related variations discussed above. To quantify agreement, we use the Spearman rank correlation; additionally, the average αB values are reported relative to the median of the averages, which is needed because SuperSynthIA and HMIP disagree on the field strength of quiet Sun.

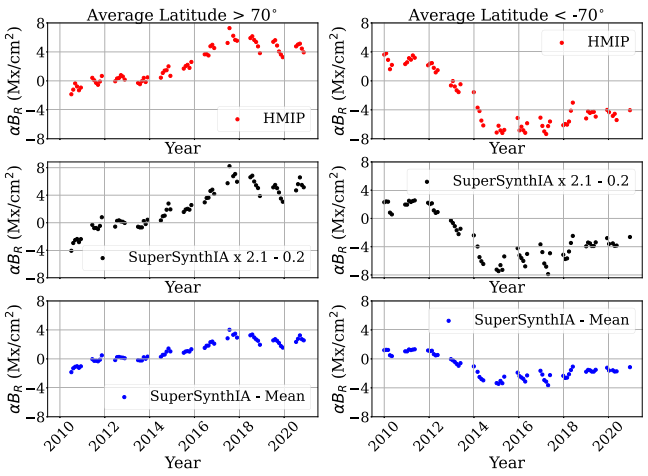

Figure 22. Long-term polar behavior of SuperSynthIA over a decade. We plot the north (left) and south (right) poles. We define polar regions with a threshold of 70° and show only observations in the most visible half of the year for each pole. From top to bottom: HMIP, SuperSynthIA when optimally linearly scaled to match HMIP, and the raw predictions of SuperSynthIA. There is close agreement between SuperSynthIA and HMIP. The slight miniature trends within each year are likely due to varying pixel populations as the viewing angle changes over the year.

SuperSynthIA tracks HMIP well over long periods of time. All Spearman rank correlations are $\geqslant$0.7, even when using all pixels, including limb pixels where SuperSynthIA and HMIP are already known to differ. A more natural comparison is between pixels where SuperSynthIA and HMIP are most likely to agree, namely, strong polarization signals (by proxy, where `conf_disambig` $\geqslant 60$) and disk-center ($\mu > 0.5$) pixels; limiting the input to these pixels, the Spearman rank correlations are far higher: 0.95 for αB and 0.87 for αB_R. Moreover, while a correction factor is indeed required to align SuperSynthIA and HMIP for αB, no such factor is required from αB_R.

6.7. Vector Fields over Years

Following similar reasoning that motivated our study of trends over days, we next study trends over years. We track the reversal of polarity at the Sun's poles as described by Sun et al. (2015) and Janardhan et al. (2018). We avoid potential training overlap issues by training a model identically to our main SuperSynthIA model but using only equatorial samples. While this model has seen observations taken during the studied time period, it has not seen polar regions.

We plot the average polar-region αB_R in Figure 22, showing both the HMIP and SuperSynthIA's predictions for the north and south poles of the Sun. The output from SuperSynthIA has a strong correlation with HMIP (Pearson linear correlation coefficients near 1). When optimally linearly scaled, the predictions of SuperSynthIA align closely with those of HMIP and replicate the transition documented in HMIP: a northern pole reversal containing a sustained constant near-zero field strength from 2012 through 2014, and a clear reversal in the southern hemisphere in 2013. This trend is in agreement with Janardhan et al. (2018) and roughly agrees with Sun et al. (2015), despite the fact they used a different definition of polar latitudes (Sun, 60°; Janardhan, 55°).

7. Discussion

This paper introduces an approach, SuperSynthIA, for producing estimates of the photospheric vector magnetic field directly from HMI polarization images, using a deep learning model trained on joint SDO/HMI and Hinode/SOT-SP observations spanning many years and billions of pixels. SuperSynthIA represents a significant capability advance from the previous approach, SynthIA (Higgins et al. 2021), in that we infer the heliographic components B_R, B_ϕ, B_θ; in other words, SuperSynthIA can effectively perform the inversion and disambiguation steps of the HMI vector-field data pipeline simultaneously.

Results show that SuperSynthIA produces robust vector magnetograms that are promising for direct use. In particular, SuperSynthIA shows good characteristics in both space and time that may be critical for some downstream applications. For instance, the dramatic reduction in scan-to-scan temporal inconsistencies in the magnetic vectors' magnitudes and directions should enable more, or at least easier, use of HMI-acquired data for data-driven modeling efforts, such as studies of energetic events (i.e., flares and coronal mass ejections). Finally, the dramatic reduction in the noise in areas of weak polarization signals, the lack of azimuthal bias compared to SynthIA, and the "sharper" distributions of medium-polarization-signal areas may assist the use of HMI-acquired data for broader scientific inquiries into, e.g., plage regions and full-Sun data analysis.

Some of the improvements stem directly from setting the "answer" for SuperSynthIA to HinodeP, and specifically the new Level 2.1 disambiguated data product that includes the heliographic vector components. For instance, the reduction in noise in the weak field stems from the fact that HinodeP has a similar noise characteristic to SuperSynthIA's results. The success of SuperSynthIA should thus be credited to not only machine learning but also the extensive efforts over the past decade or more to continue the operation of Hinode and SDO, maintain near-constant capture (especially by SDO/HMI), and continue to improve and distribute the data.

On the other hand, some of the improvements do come from learning. The standard promised improvement from learning-based approaches is acceleration, as noted by virtually all works in this area (e.g., Socas-Navarro et al. 2001; Asensio Ramos & Díaz Baso 2019; Sainz Dalda et al. 2019; Liu et al. 2020 among dozens of others). We also see faster inversion as a benefit (in the case of SuperSynthIA, the speedup is roughly 2 orders of magnitude following the analysis of Higgins et al. 2021), but note two more benefits that are often overlooked. First, our method depends on only having paired data. In this paper, we have used the loose requirement of only needing paired data to train *across* instruments as in Higgins et al. (2022). However, the approach may enable the more rapid deployment of improved but more complex inversion techniques, whether for ME models such as by Griñón-Marín et al. (2021) or non-LTE methods such as by Socas-Navarro et al. (2015; as done by Socas-Navarro & Asensio Ramos 2021). As illustrated by the use of a small fraction of the total SDO/HMI pixels for training SuperSynthIA, one could imagine running a more complex inversion on a judiciously selected set of pixels and extending it to the full-disk via SuperSynthIA. The second advantage is that SuperSynthIA is learned over *multiple* pixels, and so it produces consistent results. Thus, when the input data changes slowly, the output

data tends to change slowly, leading to substantially better short-term temporal consistency.

We wish to be careful about our attribution to learning. It is tempting to ascribe the performance to *deep* learning and the particular form of the neural network architecture that was used. We believe the key feature is learning capacity, not particular forms. We experimented with other approaches on the same data set and found that several nondeep approaches (including nearest neighbors) yielded good ME inversions when applied per pixel. However, the approaches were ineffective at disambiguation apart from providing a coarse guess akin to as-radial-as-possible; this limitation could be fixed with additional context. The alternate approaches, of course, came with downsides: a large-scale nearest neighbor method, for instance, was slow even with advanced approximate search. Thus, the architecture of SuperSynthIA serves not as the *only* solution, but instead as a computationally convenient way to encode a large data set in a fast-to-evaluate function, and a way to automatically identify how much context was needed to retrieve the full heliographic magnetic field vector.

More broadly speaking, we wish to explicitly reject an incorrect view of this line of work as "replacing" existing methods and pipelines or worse, "replacing" instruments and the need for domain expertise. First, while SuperSynthIA has made tremendous progress and may be particularly well suited for some applications, the current SDO/HMI pipeline data have several clear advantages, ranging from the known issues of SuperSynthIA discussed in Section 6.1 as well as an immense amount of testing, care, and analysis over the past 13 years of use. We expect that, as SuperSynthIA is used, the community will more clearly discover SuperSynthIA's strengths and limitations, just as it discovered the strengths and limitations of the SDO/HMI pipeline. We hope that these strengths and limitations will inform the development of subsequent learning-based systems, as well as data acquisition and preparation. Second, from a purely practical perspective, SuperSynthIA fundamentally depends on both excellent systems during training and testing: SuperSynthIA's inputs come from the high-quality well-calibrated spectra produced by SDO/HMI; its target answers come from an optimization-based inversion and disambiguation approach applied to Hinode/SOT-SP data. Extensive modeling and calibration are needed to produce these data.

We found that extensive data expertise was critical from the perspective of developing SuperSynthIA and likely will be needed for deployment as well. Analysis and development required unusually close collaboration since most of the important analysis and motivation of the work comes from solar-physics specific expertise. At the same time, the learned nature of the inversions required an understanding of machine learning to diagnose issues (e.g., why median decoding produced a hemispheric bias that the training ground-truth does not have when viewed from a purely optimization lens). We are currently investigating the integration of SuperSynthIA with downstream tasks, and find this close collaboration important: SuperSynthIA produces magnetograms with different characteristics than either HMIP and HinodeP; even seemingly good properties and "improvements" can require changes to downstream applications. For instance, the increased spatial sharpness of SuperSynthIA leads to changes in spatial and temporal derivatives via finite-differences, so that windowing functions used in, e.g., plasma velocity estimations (Welsch et al. 2004; Schuck 2008) may need to be revisited by users.

We see SuperSynthIA as the start of learning-based systems that are ready for use in real downstream applications, rather than as a conclusion or final word. In this respect, SuperSynthIA provides a family of solutions for producing vector magnetograms that users may be able to adapt to their uses. For example, we demonstrate different strengths between the two approaches to estimating the heliographic vector components (either directly predicting the vector components or analytically via estimates of the disambiguated image-plane decomposition). Users not focusing on the pole or the farthest limbs may wish to use the direct approach, while explorations of limbs and poles may benefit from the analytic approach. Similarly, as better HinodeP data become available, we will release upgraded SuperSynthIA models. More broadly, we expect that other users may be able to take our techniques and develop even further improved methods.

Acknowledgments

All SDO data used are publicly available from the Joint Science Operations Center (JSOC) at Stanford University supported by NASA Contract NAS5-02139 (HMI); see http://jsoc.stanford.edu/; which also supported J.T.H. and Y.L. Hinode is a Japanese mission developed and launched by ISAS/JAXA, with NAOJ as domestic partner and NASA and STFC (UK) as international partners. It is operated by these agencies in cooperation with ESA and NSC (Norway). After publication, data and models used will be archived and given Digital Object Identifiers (DOIs), and D.F. will be able to provide runs-on-request. This work was started under and partly supported by NASA grant 80NSSC20K0600 for the NASA Heliophysics DRIVE Science Center (SOLSTICE) at the University of Michigan. K.D.L. and G.B. acknowledge support from NASA/LWS-SC 80NSSC22K0892, and Lockheed-Martin Space Systems contract No. 4103056734 for Solar-B FPP Phase E. D.F., R.W., K.D.L., and G.B. acknowledge support from NASA award 80NSSC22K0646. D.F. and R.W. thank Michigan CSE Division and the University of Michigan Advanced Research Computing for continued HPC support. The authors thank David Orozco Suárez for thoughtful feedback that improved the paper.

Appendix

We now include a number of extra figures that help more fully explain evaluations or provide additional analyses and results that support claims in the main paper.

Appendix A
Viewing Angle Dependence

Figure 18 of the main paper establishes the existence of a center-to-limb bias over all pixels (although we stress that there is no bias when one examines active region pixels). For completeness, we show how errors vary in Figure 23 for αB as well as the heliographic components αB_R, αB_ϕ, αB_θ. In principle, a method could avoid a viewing angle artifact by simply having a consistently bad estimate of field properties (e.g., $\alpha B = 0\,\mathrm{Mx\,cm^{-2}}$).

Given a sequence of predictions $\{\hat{y}_i\}_{i=1}^N$ and ground-truths $\{y_i\}_{i=1}^N$, we compute the signed error $\epsilon_i = \hat{y}_i - y_i$. A positive

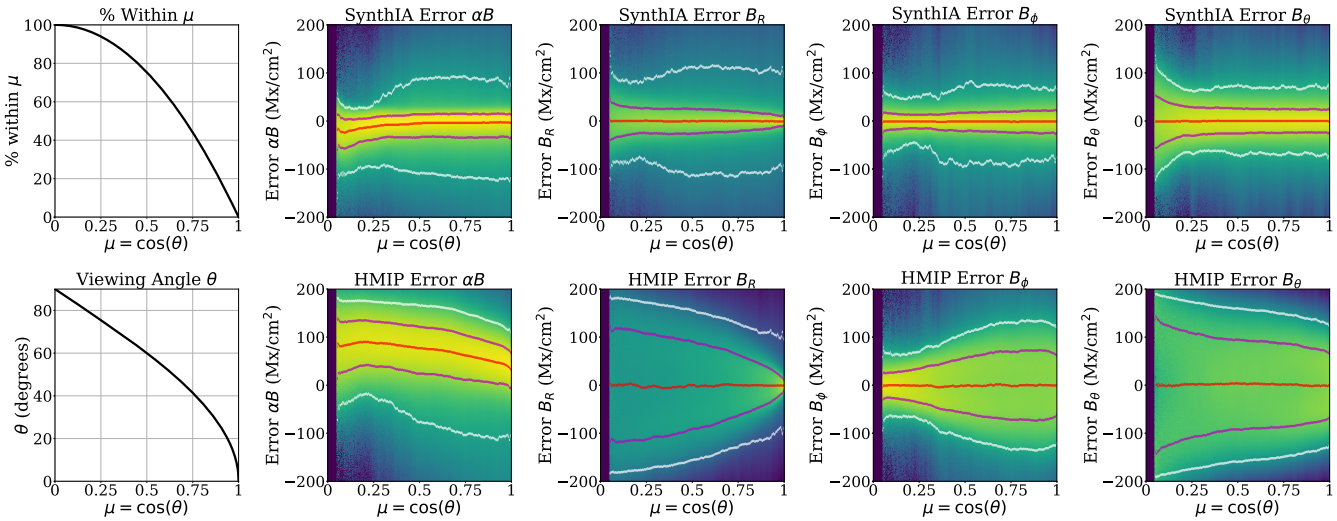

Figure 23. Error with respect to HinodeP as a function of viewing angle for equatorial scans over all pixels. In each plot, the x-axis represents the viewing angle $\mu = \cos(\theta)$, where 1 is at disk center, and 0 is at the limbs. (Left column): we plot the percent of data as a function of μ (top) as well as the angle (bottom) to help provide context for the error histograms on the right. (Right four columns): each panel shows a histogram of errors as a function of viewing angle μ. On top of the log-density (minimum maximum log-density), we show the median in red, 66% confidence intervals in magenta, and 95% confidence intervals in white. The top row shows errors for SuperSynthIA for αB, B_r, B_ϕ, and B_θ; the bottom row shows the same for HMIP. HMIP's behavior varies substantially more as a function of μ compared to SuperSynthIA due to the noise artifact in low-signal regions: αB becomes a higher overestimate on the limbs, and both the B_r and B_θ errors spread out considerably. Note that the artifact in HMIP does not affect pixels with good measurement (i.e., `conf_disambig` $\geqslant 60$).

signed error indicates an overestimate, and a negative signed error is an underestimate. We plot the distribution of errors ϵ_i conditioned on μ for both HMIP and SuperSynthIA as a function of μ in Figure 23 for equatorial pixels. Results are similar for polar regions. HMIP's errors show substantial variability as a function of μ, and especially more so than SuperSynthIA. HMIP's tendency to substantially overestimate quiet strength pixels can be seen by the strong tendency toward overestimates closer to the limb: the median error rises to nearly 100 Mx cm^{-2} at the limbs. This translates to substantial variability as a function of μ for the heliographic components: errors become increasingly likely to be large closer to the limb. In contrast, SuperSynthIA shows substantially more stable performance as a function of viewing angle. There is some clear variability; however, for αB_R and αB_θ, the error flaring out occurs at closer to $\mu = 0.25$ or $\theta \approx 75°$, which matches where the flaring out occurs in Figure 18.

Appendix B
Short-term Temporal Consistency

In the main body of the paper, we showed observation-to-observation consistency results considering pixels that were considered confident by the HMIP disambiguation algorithm. For completeness, we also report results using all pixels. We report the results in Figure 24. SuperSynthIA continues to substantially outperform HMIP in temporal consistency for large changes, and now outperforms HMIP in temporal consistency for small changes.

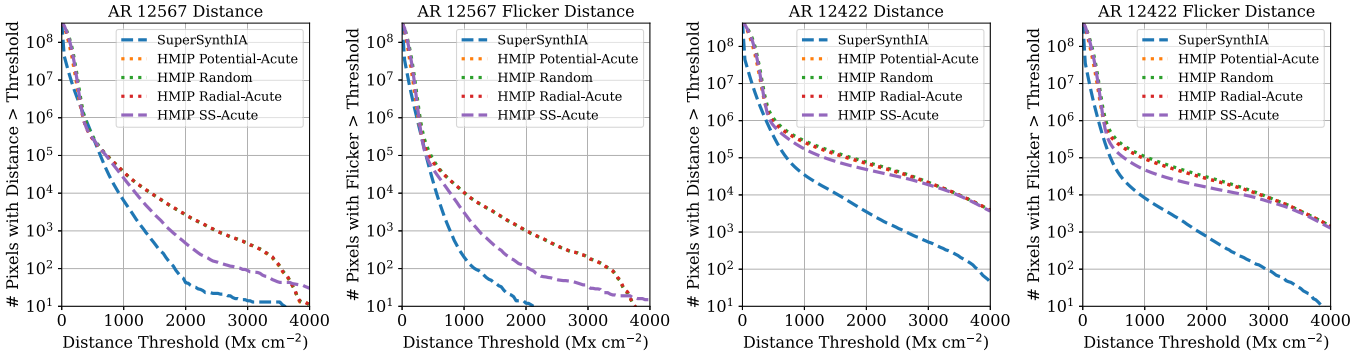

Figure 24. Same as Figure 20 using all pixels, as opposed to confident pixels. When all pixels are considered, SuperSynthIA shows an improvement at all thresholds, and not just high thresholds. The three different disambiguation methods for HMIP produce slightly different results, but are nearly identical: the regions where the disambiguation methods differ have too small of a magnetic field to have large differences. Understandably, if there is a difference, random disambiguation tends to have the most observation-to-observation inconsistencies.

Appendix C
Additional Qualitative Results

We conclude with some additional results. First, we show an additional panel of results for another cutout from the 2016 test set in Figure 25 along with a zoom-in in Figure 26. These panels show that SuperSynthIA does a good job of reproducing Hinode/SOT-SP, even when zoomed in on a $75'' \times 55''$ cutout.

Finally, for completeness, we show a comparison between SuperSynthIA and SynthIA in predicting the azimuth. One of the improvements in SuperSynthIA is mitigating preferential directions in azimuth predictions in areas with weak transverse magnetic field ($\alpha B_{\perp}$). We show four samples from the 2016 cutout test set in Figure 27. SuperSynthIA produces more realistic predictions that match HinodeP compared to SynthIA.

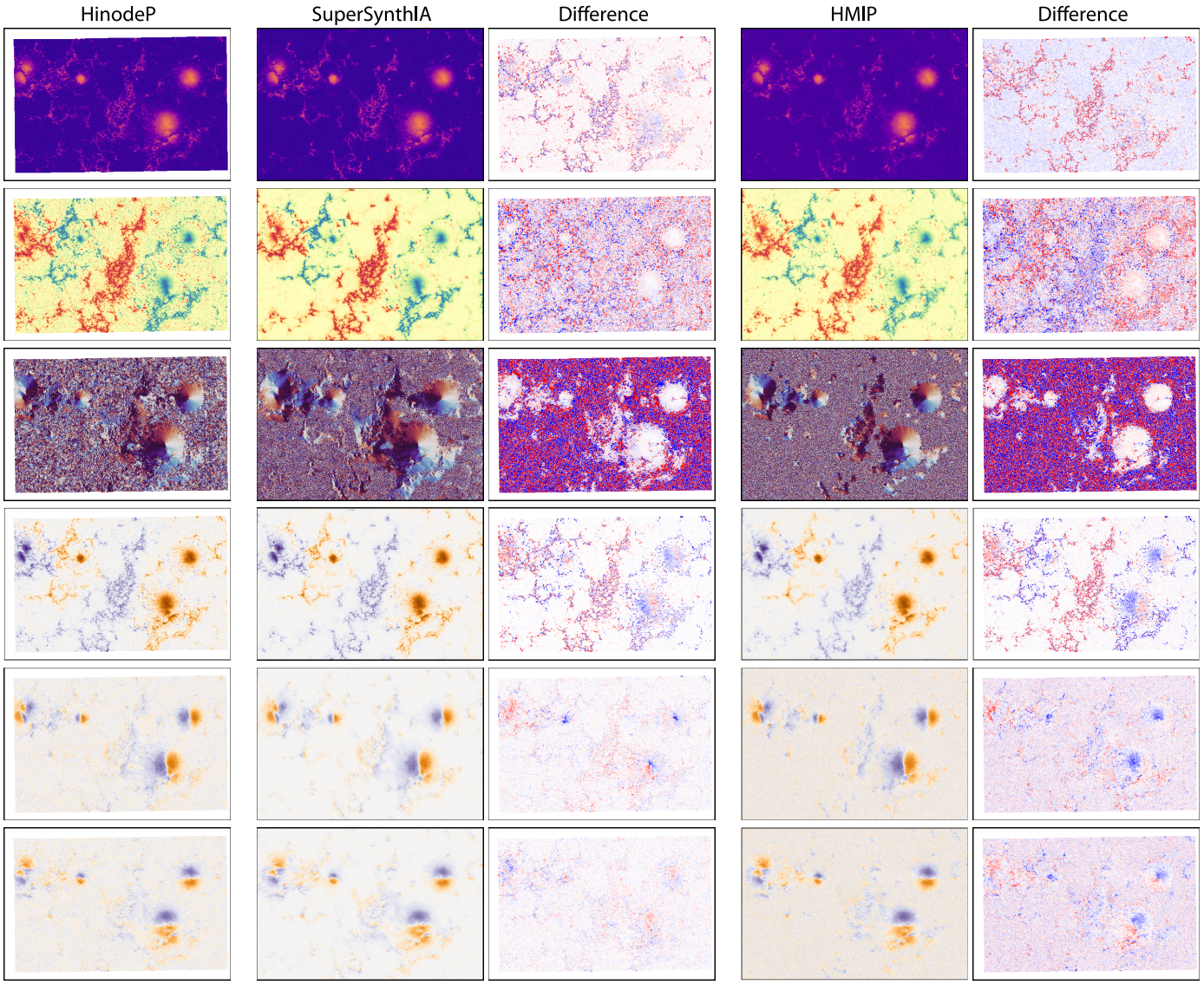

Figure 25. Additional sample of Hinode/SOT-SP, SuperSynthIA, and HMIP on 2016 August 14, 18:48 TAI. We show, in order: magnetic flux density αB, inclination γ, disambiguated azimuth ψ_{360}, and the three heliographic magnetic flux density components αB_R, αB_ϕ, and αB_θ. SuperSynthIA does a good job of imitating Hinode/SOT-SP compared to HMIP. Colormaps follow Figures 9 and 13.

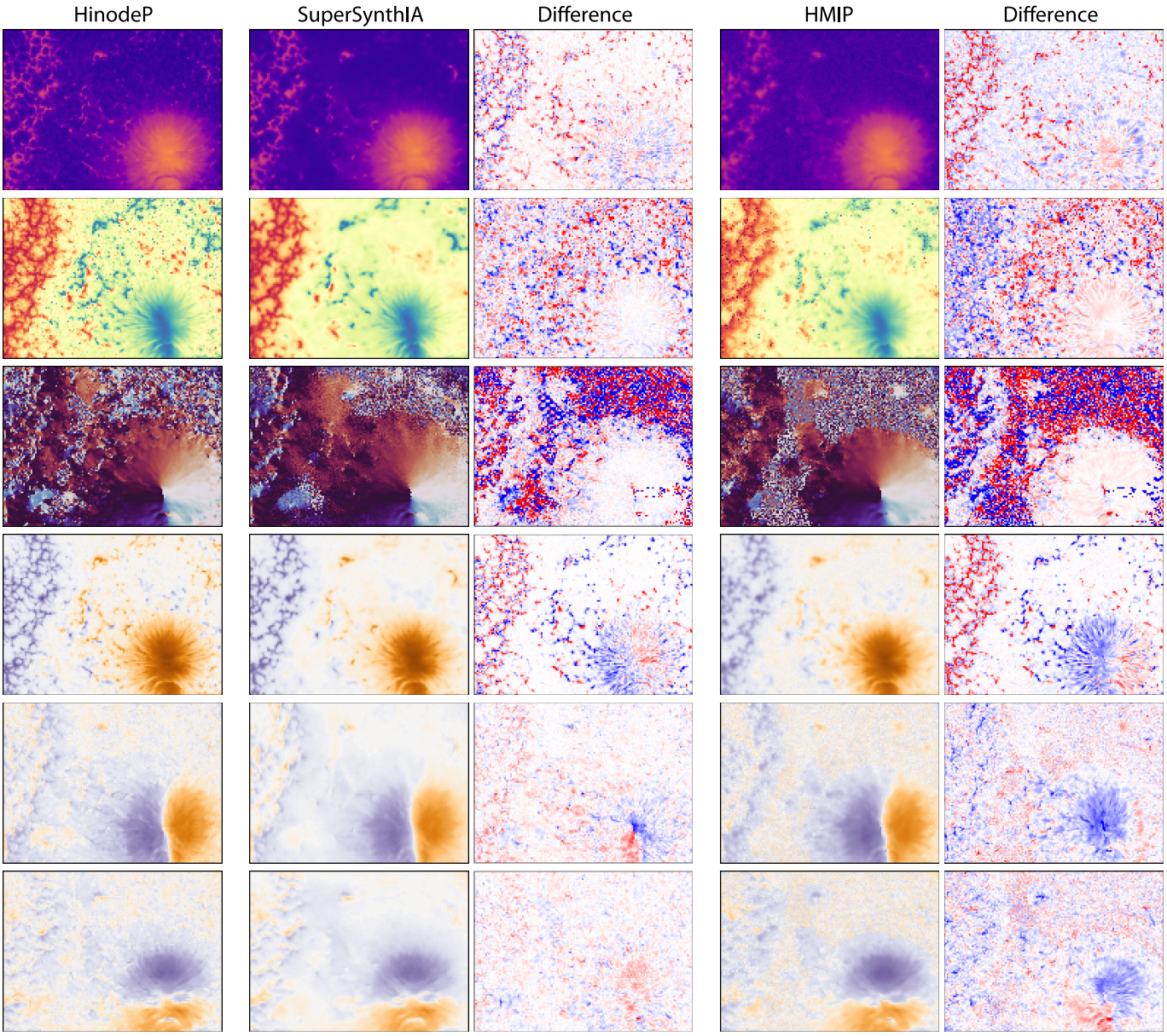

Figure 26. Additional sample of Hinode/SOT-SP, SuperSynthIA, and HMIP on 2016 August 14, 18:48 TAI, zoomed in on a $75'' \times 55''$ cutout at the large sunspot. Overall SuperSynthIA more accurately reproduces HinodeP. SuperSynthIA accurately reproduces many systematic differences between HinodeP and HMIP. Like SynthIA, SuperSynthIA reproduces HinodeP's quiet Sun value and broader networks of more highly inclined plage that were seen in Higgins et al. (2022). SuperSynthIA also extends the ME0 disambiguation into the weak regimes whereas HMIP falls back on a weak-field solution (here, the random solution). Colormaps follow Figures 9 and 13.

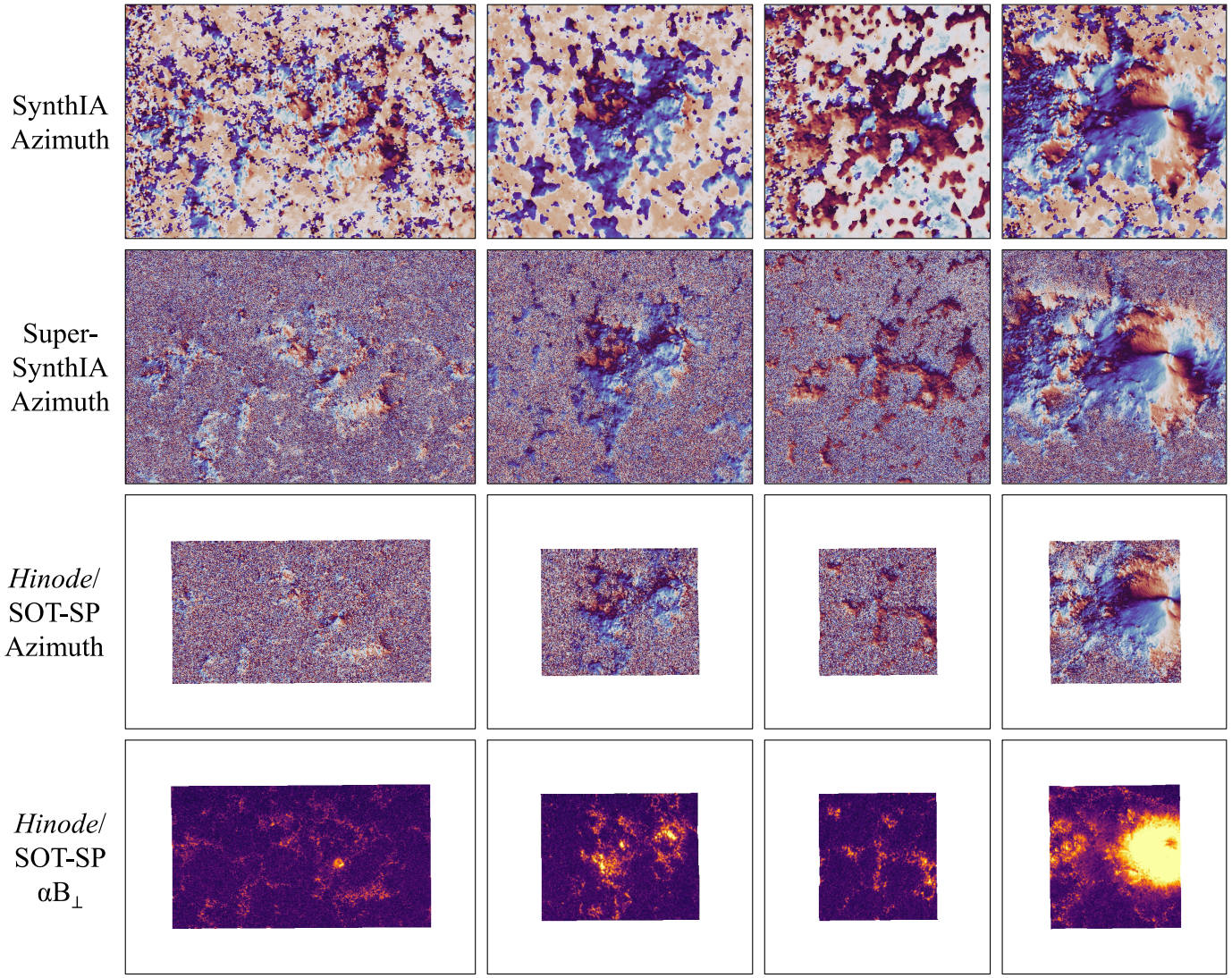

Figure 27. Comparison of Synthia with SuperSynthIA in predicting azimuth. We show SynthIA, SuperSynthIA, and Hinode/SOT-SP along with the Hinode/SOT-SP transverse magnetic field $\alpha B_{\perp}$. When $\alpha B_{\perp}$ is small, azimuth is poorly defined. Despite this, SynthIA confidently predicts a consistent magnetic field in the quiet region. On the other hand, SuperSynthIA tends to have a more uniform distribution akin to Hinode/SOT-SP. The white borders around Hinode/SOT-SP are the 100 pixel buffers placed around the scan for which there is SDO/HMI data, but no Hinode/SOT-SP data. Colormaps: azimuth, 0° 180°; $\alpha B_{\perp}$, 0 1000 Mx cm^{-2}. Examples (left to right): 2016 March 7, 12:36 TAI; 2016 January 17, 10:36 TAI; 2016 June 6, 11:00 TAI; 2016 April 13, 4:00 TAI.

ORCID iDs

Ruoyu Wang https://orcid.org/0009-0007-7942-826X
David F. Fouhey https://orcid.org/0000-0001-5028-5161
Richard E. L. Higgins https://orcid.org/0000-0002-6227-0773
Spiro K. Antiochos https://orcid.org/0000-0003-0176-4312
Graham Barnes https://orcid.org/0000-0003-3571-8728
J. Todd Hoeksema https://orcid.org/0000-0001-9130-7312
K. D. Leka https://orcid.org/0000-0003-0026-931X
Yang Liu https://orcid.org/0000-0002-0671-689X
Peter W. Schuck https://orcid.org/0000-0003-1522-4632
Tamas I. Gombosi https://orcid.org/0000-0001-9360-4951

References

Asensio Ramos, A., & Díaz Baso, C. J. 2019, A&A, 626, A102
Barnes, G., Leka, K., Schrijver, C., et al. 2016, ApJ, 829, 89
Bobra, M. G., Sun, X., Hoeksema, J. T., et al. 2014, SoPh, 289, 3549
Borrero, J., Lites, B., Lagg, A., Rezaei, R., & Rempel, M. 2014, A&A, 572, A54
Borrero, J., Tomczyk, S., Kubo, M., et al. 2011, SoPh, 273, 267
Centeno, R., Schou, J., Hayashi, K., et al. 2014, SoPh, 289, 3531
Cheung, M. C., & DeRosa, M. L. 2012, ApJ, 757, 147
Community Spectropolarimetric Analysis Center (CSAC), Lites, B. W., Garcia, A. G., et al. 2006, Level 2 Hinode/SP data, NCAR, doi:10.5065/D6JH3J8D
CSAC, C. 2006, MERLIN, https://www2.hao.ucar.edu/csac/csac-spectral-line-inversions Unlink
DeGrave, A. J., Janizek, J. D., & Lee, S.-I. 2021, NatMI, 3, 610
del Toro Iniesta, J. C., & Ruiz Cobo, B. 2016, LRSP, 13, 4
Fouhey, D. F., Gupta, A., & Hebert, M. 2013, in Proc. of the 2013 IEEE Int. Conf. on Computer Vision (Los Alamitos, CA: IEEE Computer Society Press), 3392
Fouhey, D. F., Higgins, R. E. L., Antiochos, S. K., et al. 2023, ApJS, 264, 49
Franz, M., Collados, M., Bethge, C., et al. 2016, A&A, 596, A4
Gary, G. A., & Hagyard, M. J. 1990, SoPh, 126, 21
Gary, G. A., Moore, R. L., Hagyard, M. J., & Haisch, B. M. 1987, ApJ, 314, 782
Gombosi, T. I., van der Holst, B., Manchester, W. B., & Sokolov, I. V. 2018, LRSP, 15, 1

Griñón-Marín, A. B., Pastor Yabar, A., Liu, Y., Hoeksema, J. T., & Norton, A. 2021, ApJ, 923, 84
Hartley, R. I., & Zisserman, A. 2004, Multiple View Geometry in Computer Vision (2nd ed.; Cambridge: Cambridge Univ. Press)
Harvey, J. 1969, PhD thesis, Univ. of Colorado, Boulder
Hayashi, K., Abbett, W. P., Cheung, M. C., & Fisher, G. H. 2021, ApJS, 254, 1
Higgins, R. E. L., Fouhey, D. F., Antiochos, S. K., et al. 2022, ApJS, 259, 24
Higgins, R. E. L., Fouhey, D. F., Zhang, D., et al. 2021, ApJ, 911, 130
Hoeksema, J. T., Abbett, W. P., Bercik, D. J., et al. 2020, ApJS, 250, 28
Hoeksema, J. T., Liu, Y., Hayashi, K., et al. 2014, SoPh, 289, 3483
Hurlburt, N., Slater, G., Tarbell, T., Berger, T., & Katsukawa, Y. 2009, Hinode Solar Optical Telescope Data Analysis Guide, SOT Data Guide
Ioffe, S., & Szegedy, C. 2015, arXiv:1502.03167
Jabbour, S., Fouhey, D. F., Kazerooni, E., Sjoding, M. W., & Wiens, J. 2020, in Proc. of the Machine Learning for Healthcare Conference 5, ed. F. Doshi-Velez et al. (PMLR), 750
Janardhan, P., Fujiki, K., Ingale, M., Bisoi, S. K., & Rout, D. 2018, A&A, 618, A148
Kosugi, T., Matsuzaki, K., Sakao, T., et al. 2007, SoPh, 243, 3
Kullback, S., & Leibler, R. A. 1951, Ann. Math. Stat., 22, 79
LaBonte, B. 2004, SoPh, 221, 191
Leka, K., Wagner, E. L., Griñón-Marín, A. B., Bommier, V., & Higgins, R. E. 2022, SoPh, 297, 121
Leka, K. D., Barnes, G., & Crouch, A. 2009, in ASP Conf. Ser. 415, The Second Hinode Science Meeting: Beyond Discovery-toward Understanding, ed. B. Lites et al. (San Francisco, CA: ASP), 365
Leka, K. D., Barnes, G., & Wagner, E. L. 2018, JSWSC, 8, A25
Lionello, R., Velli, M., Downs, C., Linker, J. A., & Mikić, Z. 2014, ApJ, 796, 111
Lites, B., Casini, R., Garcia, J., & Socas-Navarro, H. 2007, MmSAI, 78, 148
Liu, C. 2009, PhD thesis, Massachusetts Institute of Technology
Liu, H., Xu, Y., Wang, J., et al. 2020, ApJ, 894, 70
Liu, Y., Griñón-Marín, A. B., Hoeksema, J. T., Norton, A. A., & Sun, X. 2022, SoPh, 297, 17
Liu, Y., & Schuck, P. W. 2012, ApJ, 761, 105
Lowe, D. 2004, IJCV, 60, 91
MacQueen, J. 1967, in 5th Berkeley Symp. Math. Statist. Probability, 281
Metcalf, T. R. 1994, SoPh, 155, 235
Metcalf, T. R., Leka, K. D., Barnes, G., et al. 2006, SoPh, 237, 267
Pathak, D., Girshick, R., Dollár, P., Darrell, T., & Hariharan, B. 2017, in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (Piscataway, NJ: IEEE)
Pesnell, W. D., Thompson, B. J., & Chamberlin, P. C. 2012, in The Solar Dynamics Observatory (SDO), ed. P. Chamberlin, W. D. Pesnell, & B. Thompson (New York: Springer), 3
Pevtsov, A. A., Canfield, R. C., & Metcalf, T. R. 1994, ApJL, 425, L117
Pevtsov, A. A., Liu, Y., Virtanen, I., et al. 2021, JSWSC, 11, 14
Rachkovsky, D. N. 1962, IzKry, 28, 259
Ronneberger, O., Fischer, P., & Brox, T. 2015, in Int. Conf. on Medical Image Computing and Computer-assisted Intervention, ed. N. Navab (Berlin: Springer),
Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. 2011, in Int. Conf. on Computer Vision (Piscataway, NJ: IEEE), 2564
Sainz Dalda, A. 2017, ApJ, 851, 111
Sainz Dalda, A., de la Cruz Rodríguez, J., De Pontieu, B., & Gošić, M. 2019, ApJL, 875, L18
Scharstein, D., & Szeliski, R. 2002, IJCV, 47, 7
Schou, J., Scherrer, P. H., Bush, R. I., et al. 2012, SoPh, 275, 229
Schuck, P. W. 2008, ApJ, 683, 1134
Schuck, P. W., & Antiochos, S. 2019, ApJ, 882, 151
Schuck, P. W., Antiochos, S. K., Leka, K., & Barnes, G. 2016, ApJ, 823, 101
Sigwarth, M., Balasubramaniam, K. S., Knölker, M., & Schmidt, W. 1999, A&A, 349, 941
Snavely, N., Seitz, S., & Szeliski, R. 2006, ACM Transactions on Graphics, 25, 835
Socas-Navarro, H., Ariste, A. L., & Lites, B. 2001, ApJ, 553, 949
Socas-Navarro, H., & Asensio Ramos, A. 2021, arXiv:2101.11445
Socas-Navarro, H., de la Cruz Rodriguez, J., Ramos, A. A., Bueno, J. T., & Cobo, B. R. 2015, A&A, 577, A7
Sun, X., Hoeksema, J. T., Liu, Y., & Zhao, J. 2015, ApJ, 798, 114
Tsuneta, S., Ichimoto, K., Katsukawa, Y., et al. 2008, SoPh, 249, 167
Unno, W. 1956, PASJ, 8, 108
Welsch, B., Fisher, G., Abbett, W., & Regnier, S. 2004, ApJ, 610, 1148
Westendorp Plaza, C., del Toro Iniesta, J. C., Ruiz Cobo, B., et al. 1998, ApJ, 494, 453
Wiegelmann, T., & Sakurai, T. 2021, LRSP, 18, 1